Methodological aspects and technical approaches in measuring subjective well-being

Università degli Studi di Firenze

WORKING PAPER

Filomena Maggino

The state of the art

in indicators construction

To go deeper (A)

Methodological aspects and technical approaches in measuring subjective well-being

Many international initiatives and events reveal the increasing attention on individual perception of living conditions and quality of life, also in consideration of the limits of objective information in describing social well-being.

All this requires a deep reflection and particular attention from statisticians on the necessity of a correct scientific approach to measurement and analysis of subjective data.

In fact, measuring subjective characteristics and creating subjective data require a particular attention and expertise methodology.

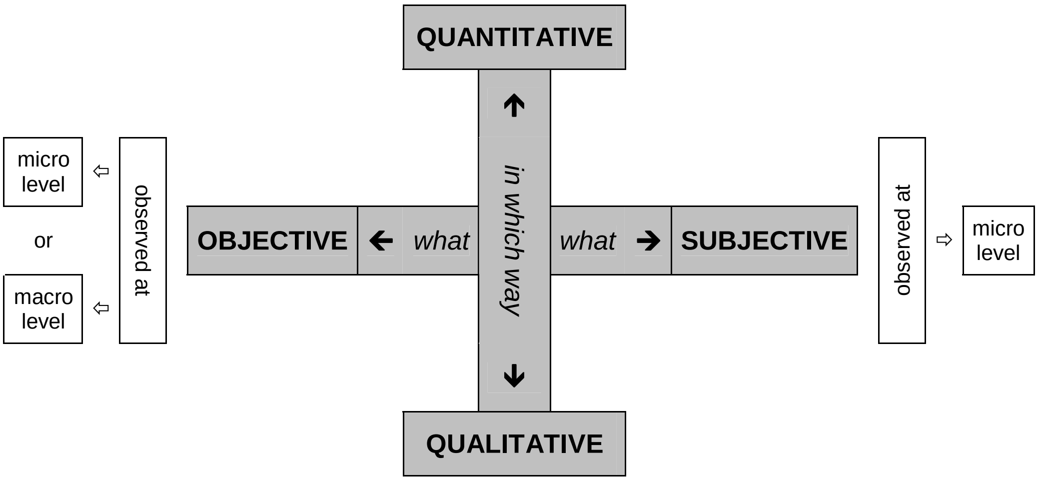
This work aims at unravelling some important methodological aspects and issues that should be considered in measuring subjective characteristics and creating subjective data.

Table of Contents

1. Objective and subjective components

Sometimes, the distinction between objective and subjective is considered equivalent to the distinction between quantitative and qualitative. Of course, this is not correct. In our perspective, we can summarize the two dyads as follows:

- “objective – subjective” refers to what we are going to observe
- “quantitative – qualitative” refers to the methodological approach applied in order to observe the previous dimensions



1.1 “Objective” and “subjective”: towards shared definitions

The necessity to study and comprehend facts through the observations of different components with reference to two different perspectives of observation, traditionally classified in terms of objective and subjective components is felt in many research fields concerning social phenomena – from economics to education).

The identification of the two aspects – objective and subjective – represents in itself a reduction of the reality. Even if the reduction is needed for measuring reasons, it should not degenerate into a contraposition between two different “realities.” The reality will be inevitably distorted by contemplating just one of the two aspects.

Before defining the two components, it could helpful trying to clarify here the meaning of “**objective**” and “**subjective**” adjectives consistently to different concepts:

A. Conceptual framework defined in order to observe a reality. In this case it is difficult to identify an objective model since the conceptual framework is always yielded by a “subjective” hypothesis and view of the world made by the researcher. Concerning this, as Michalos (1992) noticed, many models defined to observe a reality are only apparently neutral. Actually, the conceptual model is represent only a “small window” through which it is possible to see only some facets of the reality (*reductionism*); in this sense, the view is politically and socially distorted and can condition knowledge, evaluations, choices, actions, and policies.
B. Method of measurement and analysis, in this case the adjectives refer to the adopted methodologies to study the characteristics defined in the ambit of the conceptual framework: the researcher should pursue objective methodologies.
C. Observed/measured characteristics, in this case the adjectives refer to the kind of information, which has been defined in the ambit of the conceptual framework and subsequently objectively measured and analyzed. In order to make the distinction between objective and subjective characteristics more clear from the operating point of view, we can refer to the source – called *unit* – on which the characteristic of interest is measured. The units can be represented by individuals, institutions, social groups, services, administrative areas, geographical areas, nations, and so on. Consequently, we can distinguish between:
   - *objective information,* collected by observing reality
   - *subjective information*, collected only from individuals.

1.1.1 Objective components

In synthetic terms, objective components refer to the conditions in which each individual lives (health, working conditions, environmental situations, and so on). They can find different definition according to two major perspectives:

- micro-level, referring and taking into account the individual level
- macro-level, concerning and taking into account economic, demographic, geographical, administrative or social level.

Micro-level

Among the objective characteristics observed at individual level, we can mention:

- *demographic and socio-economic characteristics* (sex, age, civil status, household, educational qualification, professional condition, income, birthplace, residence, domicile, geographical/social mobility, etc.);
- *life style* that can be defined by
  - activities (work, hobby, vacation, volunteering, sport, shopping, etc.),
  - engagements (familiar, working, social, etc.),
  - habits (schedule, using of public transport and of means of communication, diet, etc.);
- *observable knowledge and skills*;
- *observable behaviours*, past and present (maybe related to the future ones).

One of the notions that can help in differentiating generic individual information from subjective information is that the latter can be observed only by/from the subject his/herself, in other words does not admit *proxy* person.

With reference to quality of life, the *objective components at micro level* refer mainly to *individual living conditions*, material resources, standards of living, working conditions and status, state of health, individual status, social relationships, freedom to choose one's lifestyle. Objective indicators allow each aspect of living conditions to be evaluated. Their specificity is in the possibility to define and recognize external objective references. In other words, they are *verifiable*.

Macro-level

It is difficult to make an inventory of all possible objective characteristics definable and observable at macro level because they are different depending on the observed and studied field. Examples can be represented by aspects concerning environmental conditions, observable social, economic and health contexts (economic production, literacy rates, life expectancy, natural and urban environmental indices, political indices, and so on).

	level				
Objective	micro	Demographic and socio-economic characteristics	- sex - age - civil/marital status - household	- educational qualification - occupation - birthplace / residence / domicile	
		Observable acquired knowledge	- skills - cognition	- know-how - competences	
		Individual living conditions and resources	- standards of living - financial resources (income) - housing	- working and professional conditions and status - state of health	
		Social capital	- social relationships - freedom to choose one's lifestyle - geographical mobility		
		Observable behaviours and life style	- activities (work, hobby, vacation, volunteering, sport, shopping, etc.)		
			- engagements (familiar, working, social, etc.)		
			- habits (schedule, using of public transport and of means of communication, diet, etc.)		
			- public life (participation, voting, etc)		
	macro	Structure of societies	Social conditions	Social exclusion	Disparities, equalities/inequalities, opportunities
				Social inclusion	Informal networks, associations and organisations and role of societal institutions
				Social mobility	
			Political setting	Human rights, democracy, freedom of information, etc.	
			Institutional setting	Educational system	
				Health system	
				Energy system	
			Economical setting	Income distribution, etc.	
			Decisional and institutional processes		
		Environmental conditions			

1.1.2 Subjective components

Traditionally "subjective characteristics" can be distinguished in three content areas (Nunnally, 1978):

- **abilities**, that concern the capacity in performing different tasks (*performance*, that is evaluated with reference to specified criteria); the abilities can be intellectual (usually thought of as those forms of abilities that are important for scholarly accomplishment and scientific work) or special (usually thought to be important for mechanical skills, artistic pursuits, and physical adroitness); among the abilities we can mention the verbal comprehension and fluency, the numerical facility, the reasoning (deductive and inductive), the ability to seeing relationships, the memory (rote, visual, meaningful, etc.), the special orientation, the perceptual speed;
- **personality traits**, that can be defined as the psychological characteristics that determine the organizational principles and that reflects the way through which an individual reacts to the environment (*locus of control*, ego, introversion, self-esteem, identification, etc.); in this perspective, some overlapping categories can be identified:
  - *social traits*, represented by the characteristic behaviour of individuals with respect to other people; typical social traits are honesty, gregariousness, shyness, dominance, humour, social responsibility, religiosity, charity;
  - *motives*, concerning individual characteristics aimed at reaching a certain goal and satisfying personal nonbiological "needs" and "drives" (affiliation, aggression, achievement, and hostility)[1];
  - *personal conceptions*, concerning the way in which the individual interacts with the social and material environment; i.e., a subject can (a) view other people as basically friendly or hostile, (b) believe that getting ahead in life depends more on luck, (c) believe important to plan personal goals on a long-range; etc.;
  - *adjustment*, concerning the relative freedom from emotional distress and/or socially disruptive

[1] Concerning this, we can mention che Abraham H. Maslow in 1954 in his work *Motivation and Personality* defined hierarchy of needs; Maslow postulated that needs are arranged in a hierarchy in terms of their potency. Although all needs are instinctive, some are more powerful than others. The lower the need is in the pyramid, the more powerful it is. The higher the need is in the pyramid, the weaker and more distinctly human it is. The lower, or basic, needs on the pyramid are similar to those possessed by non-human animals, but only humans possess the higher needs.

behaviour; this trait is strongly connected to the others (i.e., a hostile social trait makes the individual less adjustable);
- *<u>personality dynamics</u>*, that consist of organizational principles whereby the above four types of traits are "put together" (i.e., the identification with various role models); these principles help in explaining the articulation of a unique person;

- **<u>sentiments</u>**, generic terms referring to:
  - *<u>interests</u>*, concerning the preferences for particular activities;
  - *<u>values</u>*, concerning preferences for "life goals" and "ways of life"; actually, the term "value" refers to a wide range of contents, from intellectual aspects of life to more abstracts values regarding goals of self-attainment;
  - *<u>attitudes</u>*, concerning feelings about particular objects; traditionally, attitudes are defined as composed by three components:
    - <u>cognitive</u> (beliefs), important component even though not easy to be defined, concerning the way whereby the individual judges the social and material environment (**evaluations**); so, it refers also to the **opinions** that an individual has with reference to particular objects (physical objects, type of people, politics, social institutions, policies, etc.);
    - <u>affective</u>, reflecting the feelings, the evaluations, the emotions, the perceptions and the self-descriptions of an individual with reference to particular objects (i.e., professional role); this component can include also the dimensions of **satisfaction** and **well-being** for the dimensions of individual life (job, study, family, relationships, etc.) and **emotional states** (i.e., happiness);
    - <u>behavioural</u> (actual actions), reflecting the behavioural tendencies of an individual with reference to a certain object, the *<u>intentions</u>* can be included in this component, thought as actions or behaviours that the individual plans and will execute in the future.

Of course, the scheme is not exhaustive and the different identified components for each area can overlap one another.[2]

<table>
<tr><td></td><td>level</td><td colspan="4">components</td></tr>
<tr><td rowspan="8">subjective</td><td rowspan="8">micro</td><td rowspan="2">abilities / capacities</td><td>intellectual</td><td>- verbal comprehension and fluency
- numerical facility
- reasoning (deductive and inductive)
- ability to seeing relationships</td><td>- memory (rote, visual, meaningful, etc.)
- special orientation
- perceptual speed</td></tr>
<tr><td>special</td><td>- mechanical skills
- artistic pursuits</td><td>- physical adroitness</td></tr>
<tr><td colspan="2">personality traits</td><td>- social traits
- motives
- personal conceptions</td><td>- adjustment
- personality dynamics</td></tr>
<tr><td rowspan="5">sentiments</td><td colspan="3">Interests and preference</td></tr>
<tr><td colspan="3">Values</td></tr>
<tr><td rowspan="3">Attitudes</td><td colspan="2">cognitive → evaluations (beliefs, evaluations opinions)</td></tr>
<tr><td colspan="2">affective → perceptions (satisfaction and emotional states – i.e., happiness)</td></tr>
<tr><td colspan="2">behavioural intentions</td></tr>
</table>

[2] In the ambit of the Multiattribute Evaluation approach, a model (Ajzen & Fishbeim, 1980; Fishbein & Ajzen, 1975) has been defined defining attitudes in symbolic terms:

$$A_o = \sum_{i=1}^{n} b_i e_i$$

dove

A_o	attitude *A* towards object *O*
b_i	strength of the subject's trust as regards attribute *i*
e_i	evaluation of attribute *i*
n	number of attributes

Analogously, a model has been defined concerning the intentions:

$$B \approx BI = W_1(A_b) + W_2(SN)$$

dove

B	behavior
BI	intention concerning a behavior
A_b	attitude towards the accomplishment of the behavior *B*
SN	subjective norm /social influence
W_1 e W_2	empirical weights representing the relative influence of the components.

According to this model, attitudes and social influences do not directly affect individual behavior but act in individual intention.

With reference to well-being, subjective components refer to and concern opinions, evaluations, feelings, perceptions, attitudes, desires, values, and motivations related to each individual life as a whole or in different specific contexts. Contrarily to the objective characteristics, no explicit standard is defined and no external reference can be defined in observing the subjective component.
It can be assessed by individuals' or groups' responses to questions about happiness, life satisfaction, utility, or benefit. Subjective indicators aim at measuring and quantifying individual components involving different elements – as conscience, cognition, emotion, attitude, and opinion – that are related to contingent and mutable situations. Even if it is difficult to assess its measurement, social policies and programmes need more and more data concerning this component in order to complement social, economic, and health factors, such as the degree to which a perceived need is being met and the importance of that 'perceived need' to one's overall quality of life. In their review on quality of life measures, Hughes and Wang (1996) reported a classification of the possible subjective well-being indicators: satisfaction about different aspects of life, sentiments, life perceptions, values and personal aspirations, self-concept, general sense of well-being, happiness and self-esteem. The elements to be considered in planning a survey oriented to measuring subjective quality of life make indispensable an interdisciplinary approach, the only one able to consider and to understand the different levels at which each individual react to the submitted question. The different levels involve personality, values, interests, motivations, intellectual and expressive dispositions, memory, experiences, social attitudes as a member of a limited group or of a community, and so on.

Measuring techniques

In order to measure subjective characteristics, different approaches can be identified and combined in various practical and functional ways, producing quantitative or qualitative information.

- Performance measures: this approach is appropriate for measuring abilities; the measure is represented by the outcome obtained through the execution of an assigned task; this outcome is evaluated with reference to specified criteria of success; this allows intra-individual and inter-individual comparisons to be evaluated;
- Inventory measures: this approach is appropriate for measuring personality, values, interests; the measure is represented by the individual answer to a certain number of submitted "stimuli"; the inventories can be distinguished in two kinds:
  - *self inventory*: in this case the subject is asked to use the submitted stimuli in order to describe him/herself;
  - *inventory*: in this case the subject is asked to use the submitted stimuli not to describe him/herself but to describe another hypothetical individual's behaviour.
- Self-reported measures: this approach is appropriate for measuring attitudes, opinions and abilities; the measure is represented by the subject's answer – expressed in terms of agreement, preference, etc. – to a particular statement referring to the characteristic to be measured.
- Observational methods: in this case the measure is represented by the result of the observation made on the subject by an external and neutral observer.
- Projective techniques: this approach is particularly appropriate for measuring social traits, motivations, adjustment, and attitudes. The measure is represented by the individual's reaction to one situation constructed but not completed – for example, a story (eventually described also by illustrations) that the subjected has to continue and/or conclude. In this way the individual's tendency to assign his/her own characteristics, more or less desirable, to other individuals is aroused (*projection*); this method were developed mainly in psychiatry and clinical psychology as diagnostic instruments;[3] it requires strong interpretative approaches and for this are named "subjective"; in fact, the interpretation of the answers is strictly connected to the researcher's experiences. Since the projective techniques cannot be standardized, they are not considered method scientifically applicable and relevant but can usefully integrate other approaches.
- Physiological measures: they refer to the relationship between subjective traits and physiological processes. However, the scientific evidence of this connection is not completely demonstrable.

The measurement of subjective characteristics presents some difficulties, produced by specific factors; in particular, the measurement of a certain characteristic can turn out to be falsified because of:

a. *individual factors* as:
   - *social desirability*, that is the individuals' tendency to reply and react according to criteria that the individuals consider acceptable in the community (socially desirable); probably, this factor comes from the combination of different components (adjustment, knowledge, etc.) and to some extent can

[3] Wellknown techniques in this ambit are (i) *Rorschach test*, aimed at examining the personality characteristics and emotional functioning through inkblots, and (ii) *thematic apperception test*, known as the *picture interpretation technique* because it uses a standard series of provocative and ambiguous pictures about which, the subject must tell a story.

be controlled, for example, by assuring anonymous answers,
- *response set*, that is the individual's tendency to reply in a stereotyped and/or systematic way apart from the measured characteristic (*acquiescence response set*),
- *researcher's attitude*, that does not prove to be adequate and "objective" maybe because of a lacking in preparation,

b. *semantic factors* that can provoke interpretative discordance between subject and researcher,
c. *situational factors* when the observation occurs in different situations (presence or not of other persons) o in different context (at home, at work, in the street, etc.).

1.1.3 Objective and subjective components of well-being

With reference to social well-being, the two components can be articulated more minutely. Schultz (2000) propose to classify the components along a continuum ranging from "more objective" to "more subjective." This effort allows to identify quite clearly four groups of variables: (A) Social Structure, (B) Resources and Behaviour, defined in terms of living conditions (C) Evaluation of Living Conditions, and (D) Subjective Well-Being. By expanding the model elaborated by Schultz, the variables classification can be illustrated as follows:

Components →	STRUCTURE OF SOCIETY	SOCIAL STRUCTURE	STANDARD OF LIVING AND SOCIAL RELATIONSHIPS AND NETWORK (behavioural assessment)	EVALUATIONS OF LIVING CONDITIONS (cognitive assessment)	WELL-BEING (affective assessment)
	Social, political, institutional, economical setting	Socio-demographic characteristics	Resources and behaviour "objectively" reported/observed (concrete actions)	Beliefs and judgments	Subjective perceptions (feelings, emotions, self-descriptions, emotional states) ↓ Hopes – Fears Moods – Anomie Anxiety – Mental health
	Human rights Equality Schooling & education Health system Income distribution Longevity	Age Sex Occupation Income Household composition Marital status	Living conditions ↓ Housing Health Education Work conditions Personal environment	Importance of life domains and preferences for "life goals" and "ways of life" Perceived need fulfilment	Well-being ↓ Satisfaction concerning life domains Happiness
Description of →	*Country's health and wealth*	*Social stratification (e.g. occupational prestige)*	*Characteristics of society (social/economic/political system)*	*Quality of society (social/economic/ political system)*	*Subjective dispositions*

1.2 Quantitative and qualitative approaches

The previously defined conceptual framework allows characteristics to be defined "subjective" and consequently measured. In this perspective, different methodological approached can be identified, in order to observe phenomena consistently. The approaches can be broadly classified into "quantitative" and "qualitative". The two perspectives differ in many ways and often these differences have been the source of considerable debates and divisions between different researchers.
In a practical sense, there are some key differences between qualitative and quantitative research

		Qualitative	**Quantitative**
Aim		Complete and detailed description.	Constructing statistical models in order to explain what is observed. Classifying and counting features.
To be adopted when		Recommended during earlier phases of research projects	Recommended during latter phases of research projects.
Hypothesis		Exploratory perspective and purpose (i.e., hypothesis-generating).	Confirmatory perspective: testing hypotheses.
Prior knowledge		Researcher knows roughly in advance what he/she is looking for.	Researcher knows clearly in advance what he/she is looking for.
Study design		The design emerges as the study unfolds.	All aspects of the study are carefully designed before data collection.
Researcher role The basic underlying assumptions guide and sequence the types of data collection methods employed.		Researcher - is the data-gathering instrument - studies by participating and/or being immersed in a research situation - tends to become subjectively immersed in the subject matter This means that the possibility of the researcher taking a 'neutral' position is more problematic, in practical and/or philosophical terms. Hence qualitative researchers should reflect on their role in the research and analytical process.	Researcher - uses tools, such as questionnaires or equipment to collect numerical data - is ideally an objective observer who neither participates in nor influences what is being studied - tends to remain objectively separated from the subject matter.
Measures quality		Qualitative perspective aims at establishing content validity.	Quantitative perspective aims at establishing reliable and precise measures through focused hypotheses, measurement tools and applied mathematics.
Data collection		Different approaches, typically: - participant observation, - non-participant observation, - field notes, - reflexive journals, - structured interview, - group discussion - unstructured interview, - analysis of various texts, pictures, documents and other materials.	Models and hypotheses lead to development of instruments and methods for measurement.
Data	**form**	Words, pictures or objects.	Numbers, codes, statistics.
	characteristics	More 'rich', time consuming, and less able to be generalized.	More efficient, able to test hypotheses, but may miss contextual detail.
	interpretation	Subjective - individuals' interpretation of events is important, e.g., uses participant observation, in-depth interviews etc.	Objective – seeks precise measurement & analysis of target concepts, e.g., uses surveys, questionnaires etc.
	analysis	Data are - categorized into patterns as the primary basis for organizing and reporting results - analyzed through observer ideas and impressions These can be the final conclusion of the analysis, or can be subsequently analysed through quantitative approached. An example of quantitative techniques using qualitative data is textual data analysis, or content analysis.	In quantitative research, data analysis represents an important moment of the research process and is based upon statistics.
Sampling		Sample size is usually small and the sampling procedure is typically not probabilistic. Cases can be selected according to certain characteristics or contextual information.	Sample size is usually large and the sampling procedure is typically probabilistic.

More recent tendencies in social science research are aimed at adopting eclectic approaches and to use a variety of both quantitative and qualitative methods. The often-called mixed methods research has become increasingly accepted and common. It has been argued that rather than one approach being definitively more conclusive than the other; they are better understood as different practices.

Quantitative methods might be used with a global qualitative frame while qualitative methods might be used to understand the meaning of the numbers produced by quantitative methods. Using quantitative methods make possible to give precise and testable expression to qualitative ideas. This combination of quantitative and qualitative data gathering is often referred to as mixed-methods research.

2. Theoretical aspects of the measurement process

2.1 Reference theory of measurement

The reference-theory of measurement defines the theoretical characteristics that make the measurement "scientific" by defining the concept of *measurement error*. Consequently, the reference-theory allows us to identify models aimed, in particular, at assessing:

- objectivity, concerning the capacity of a procedure to measure without alteration due to external factors and to be free from effects due to the observer; this notion spreads from the procedure of measurement to the data analysis to the interpretation of the results
- precision, measured by controlling the coherence of the model of measurement; the correspondent concept of "precision" for subjective measurement is reliability
- accuracy, concerning the capacity of the procedure to measure what we intend to measure; the correspondent concept of "accuracy" for subjective measurement is validity.

A procedure of measurement that meets these requirements not only gains scientific relevance but can also be standardized.

2.2 Error in scientific measurement

As indicated above, the possibility to meet the requirements of a scientific measurement is strictly connected to the possibility to define and to identify the *error*; this represents a hypothetical component of any procedure of measurement. The observational error is the amount by which an observation differs from its expected value (Carmines & Zeller, 1992):

(error) = (measured value) – (true value)

The statistical model applied in order to evaluate the presence of error[1] in the measurement process, uses – as we will see – the concept of variability and considers two additive components:

- Random error, which may vary from observation to observation and is produced by all those uncontrolled factors that confuse and disturb the measuring; the random error is present in any measure at different amount and its effect, that can be only estimated, is completely asystematic; this means that it leads to values that can over-estimating or under-estimating the expected one;
- Systematic error (methodological error or statistical bias), which always occurs with the same value when we use the instrument in the same way; it is bias in measurement which lead to values that are systematically different from the expected one (too high or too low); in other words, a systematic error is any biasing effect which introduces error into an observational procedure and is such that it always affects the results of measuring, preventing from a correct measuring of the characteristic of interest. This error can be controlled by very carefully standardized procedures. Every scientific discipline defines and provides for procedures aimed at constructing standardized instruments.

2.2.1 Accuracy and precision

In any scientific application, accuracy and precision are closely related; in particular

- the *accuracy* is degree of conformity of a measured value to its actual (true) value, assessing accuracy requires the observation of a known process or the availability of a *reference value* (calibration);
- the *precision* is the degree to which further measurements will show the same or similar results; by determining the precision of a measurement it is possible to verify, consequently, the capacity to

[1] Since in statistics the concept of error is easily confused with the concept of residual, it could be useful differentiate them. An **error** is the amount by which an observed value differs from the corresponding expected value; the expected value is based on the whole population from which the statistical unit was randomly chosen. The errors, which are assumed to be independent from each other, are not directly observable but can be only estimated. A **residual** is an observable estimate of the unobserved error. The residuals are assumed to be not independent. So, the difference between
- the value of each case in a sample and the unobservable population average is an error,
- the value of each case in a sample and the observable sample average is a residual.

measure through a degree of distortion as low as possible; the precision is related to the concepts of *robustness* and *stability* and can be distinguished into

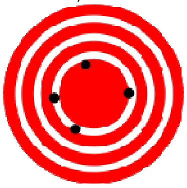
- *repeatability*, that is the variation arising when all efforts are made to keep conditions of measurement constant and by repeating during a short time period;
- *reproducibility*, that is the variation arising by using the same measurement process among different instruments and operators, and over longer time periods.

In order to explain the difference between accuracy and precision the target comparison analogy can be used. In this analogy, repeated measurements are compared to arrows that are launched at a target.

When all the arrows occupy a very narrow area, the measurements are considered precise; the size of the arrow cluster is interpreted as "precision degree".

High precision, but low accuracy

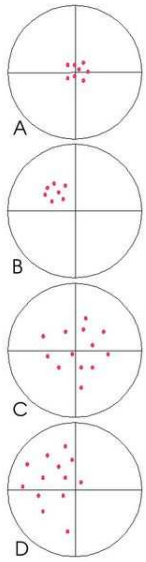
When the arrows are very close to the target's centre, the measurements are considered accurate. The distance of the arrow cluster from the centre is interpreted as the level of accuracy; in other words, the closer are the values produced by a measurement procedure to the expected value, the more accurate the procedure is considered to be.



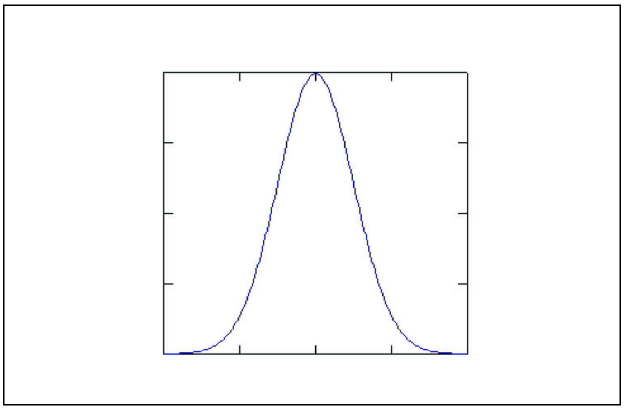
High accuracy, but low precision

da Wikipedia

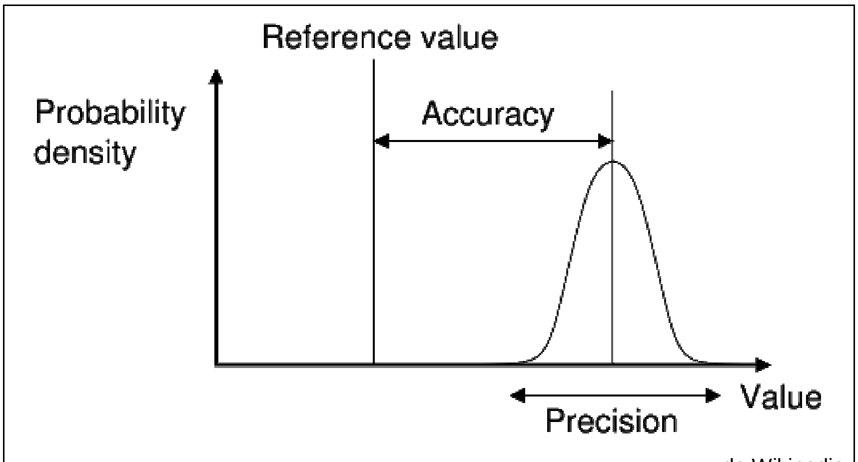
The examples besides show clusters of repeated measurements that are quite accurate and precise (A) , precise but not accurate (B), imprecise and inaccurate (C and D).



From what we have seen, we can assume that the greater the number of repetitions, the more accurate and precise the estimation of the error. By assuming repeated measures, the error is said **uncertainty**. In the frequency distribution in case of repeated measures, the value corresponding to the highest frequency is assumed to have the highest probability to be close to the true value; besides, positive errors are assumed to compensate negative errors (even if this is not actually always true). Consequently, the distribution of the values of the repeated measures is assumed to be *normal*.



The wider the range of the distribution, the greater the degree of error. Consequently, the standard deviation of the repeated measures provides for an estimation of uncertainty. In particular, the uncertainty is equal to the standard error of this distribution (standard deviation divided by the square root of the number of measurements averaged). The difference between the mean of the repeated measures and the reference value (**calibration**) represents the systematic error (*bias*).



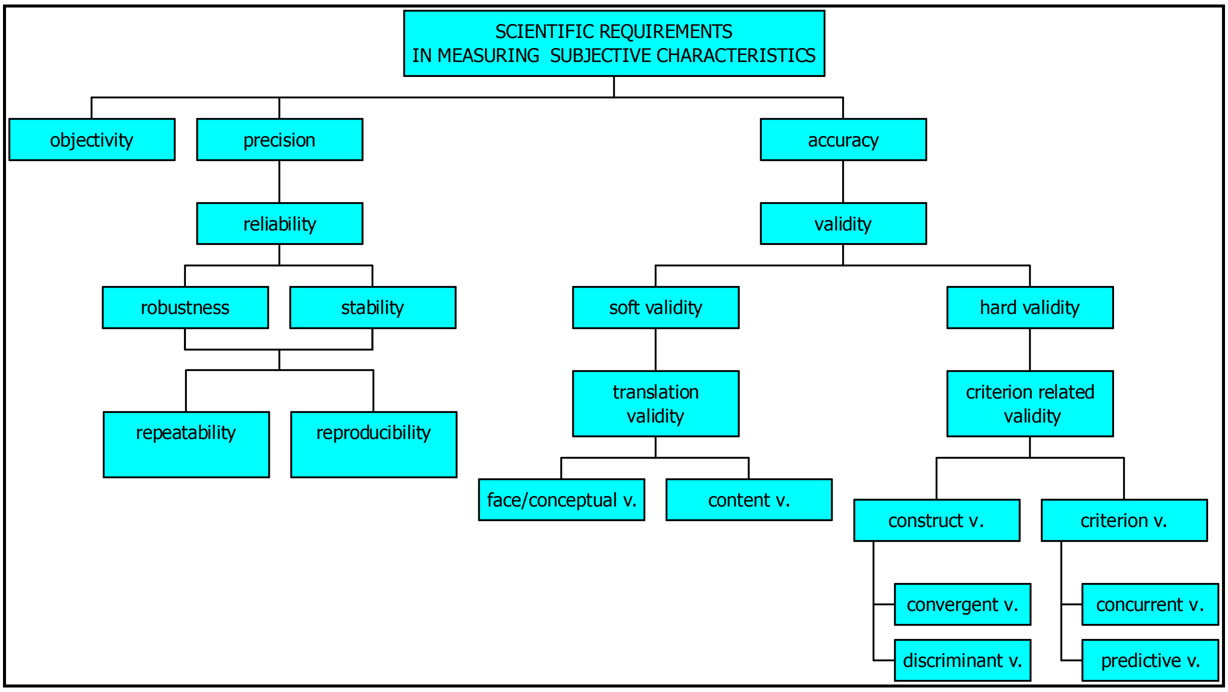
da Wikipedia

In statistics, the effect of the uncertainty of each repeated measure on the uncertainty of the whole measurement is named **propagation of uncertainty**.
Quantifying accuracy and precision is important not only in terms of correctness of measurement; in fact, the assessment of accuracy and precision allows to evaluate the costs of the presence of error for the 'economy' of the research (Biemer et al., 1991; Groves, 1989).

2.2.2 Accuracy and precision in measuring subjective characteristics

As seen above, the possibility to meet the requirements of scientific measurement is strictly connected to the possibility to define and measure the error of measurement, in its turn intimately connected to the concept of accuracy and precision. In the measurement of subjective characteristics, these concepts found an operative definition; in particular:
- the precision is measured by the degree of *reliability* that represents the consistency of the instrument of measurement (Carmines & Zeller, 1992; Ghiselli, 1964; Marradi, 1990; Netemeyer et al., 2003; Thompson, 2003; Traub, 1994);
- the accuracy is measured by the degree of *validity* that represents the tendency of the procedure of measurement to measure what is supposed to measure; in other words, validity refers to getting results that accurately reflect the concept being measured.

Reliability does not imply validity, while validity implies reliability. That means a valid measure must be reliable, but a reliable measure need not to be valid. That is a reliable measure is measuring something consistently, but not necessarily what it is supposed to be measuring.



Reliability

In the measurement of subjective measurement, in order to define the procedure allowing reliability assessment, a theory of measurement has to be identified (Laveault et al., 1994; Nunnally, 1978):

- *Classical Test theory* (*C-T*): this theory assumes the existence of a true score and a observed score; the difference between them represents the error; in order to estimate the reliability , *parallel forms* are defined; consequently, an instrument having different reliability estimations is admitted and the existence of accidental sources of error is ignored (DeVellis, 1991; Maggino, 2007; Nunnally, 1978; Spector, 1992). The reliability is mathematically defined as the ratio of the variation of the true score and the variation of the observed score or, equivalently, one minus the ratio of the variation of the *error score* and the variation of the *observed score*:

$$rho_x = \frac{\sigma_t^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

where

rho_x	reliability of the observed score, *X*
$\sigma_x^2, \sigma_t^2, \sigma_e^2$	variances on, respectively, the measured, true and error scores.

Unfortunately, there is no way to directly observe or calculate the true score, so a variety of models are defined to estimate the reliability.

- *Random Sampling theory* (*R-S*): according to this theory, the measurement of one characteristic requires a numerous group of measures (multiple measures); this theory assumes that this group is randomly drawn from a hypothetical universe of measures – concerning and completely defining the characteristic – and can be considered a statistical sample. The observed measures allow to estimating the result obtainable through the whole measures of the universe; consequently, since no explicit assumption is made regarding the stochastic process producing the observed value (Thompson, 2003), the degree of error/reliability depends on the dimension of the sample of measures (DeVellis, 1991). The estimation of reliability has to deal mainly with two problems: estimation of the score of the universe and generalization of the estimation (by the ANalysis Of the VAriance).
- *Latent Trait theory* (*L-T*): according to this theory, the measurement needs the definition of two components: the actual measurement (*indicator* or *manifest variable*) and the corresponding characteristic (not directly measurable and therefore said *latent variable*). It is assumed that 1) the responses on the indicators are the result of the individual's position on the latent variable(s), and 2) the manifest variables have nothing in common after controlling for the latent variable (local independence) (DeVellis, 1991; Maggino, 2007).

The Classical theory can be considered a special case of the Random Sampling theory in which the measures are fixed instead of having been drawn by the universe; the two theories have different definition of reliability. In fact, for the former, the reliability depends entirely on the correlations between the measures defining the characteristic to be measured (the higher is the mean of the all correlations, the higher is the

level of reliability), for the latter, the reliability depends on the dimension of the sample of measures (the higher is the number of measures, the higher is the level of reliability).
The three different theories are not completely incompatible and find an integration from the applicative point of view (Bejar, 1983; Maggino, 2007).

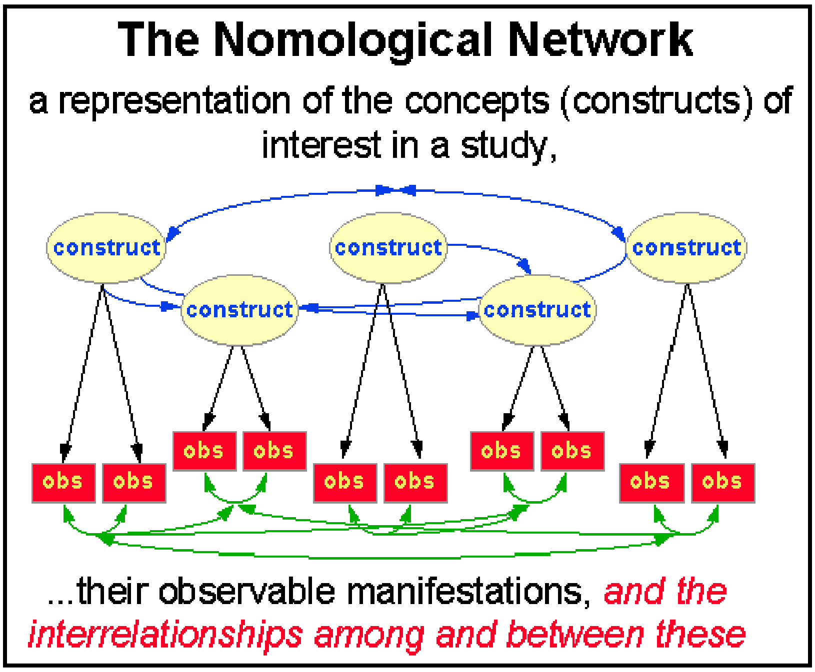
Validity

Validity represents the capacity of a measurement procedure to measure what it is supposed to measure. Different category of validity can be observed, to which a number of ways for assessing it correspond (Campbell & Russo, 2001; Carmines & Zeller, 1992; Cronbach & Meehl, 1955; DeVellis, 1991; Ghiselli, 1964; Netemeyer et al., 2003; Nunnally, 1978, Spector, 1992), as shown in the following schema:

<table>
<tr><td rowspan="4">VALIDITY</td><td rowspan="2">Soft
(translation validity)</td><td rowspan="2">The measure is a good representation of the concept that is supposed to measure</td><td>a. face validity or conceptual validity</td></tr>
<tr><td>b. content validity</td></tr>
<tr><td rowspan="2">Hard
(statistical validity)</td><td rowspan="2">The measure predicts other measures</td><td>c. criterion validity:
o concurrent validity
o predictive validity</td></tr>
<tr><td>d. construct validity:
o convergent validity
o discriminant validity</td></tr>
</table>

a. *Face Validity*, which relates to whether the measure being validated appears to be a good measure; this judgement is made on the "face" of the measure by experts.
b. *Content Validity*, which depends on a theoretical basis for assuming if the measure represents accurately the whole domain corresponding to the characteristic to be measured; it requires to be judged by experts.
c. *Criterion Validity*: which is the extent to which the measure is demonstrably related to a defined concrete criterion. It is determined by looking how much the measure correlates with another measure (criterion) known to be valid in measuring the same characteristic; significant and high correlations represent the statistical evidence of validity. When the criterion is collected:
   - at the same time as the measure being validated, the goal is to establish *concurrent validity*;
   - subsequent to the measure being validated, the goal is to establish *predictive validity* that represents the capacity of the measure to make accurate predictions on the defined criterion.
d. *Construct Validity*, which is the extent to which the measure being validated is able to measure the theoretical idea behind the characteristic under consideration; construct validity can be evaluated by statistical methods that show whether or not a supposed common factor can be shown to exist underlying several measurements. In particular, evaluation of construct validity requires examining the correlation of the measure being validated with others that are connected to costructs being or supposed to be related to the construct under consideration (Campbell & Fiske, 1959). Correlations that fit the expected pattern constitute the evidence of construct validity. Construct validity is a judgment based on the accumulation of correlations from numerous studies using the instrument being evaluated.
   Particular cases of construct validity are:
   - *Convergent Validity*, which is determined by comparing and correlating the scores obtained by the measure to be validated and the scores obtained by a measure connected to another construct, theorethically connected to the construct under consideration. Assessing convergent validity depends on the possibility to identify these relationships.
   - *Discriminant Validity*, which is specular to convergent validity; a successful evaluation of discriminant validity shows that a measure of a concept is not highly correlated with other measures designed to measure theoretically different concepts.

   The concept of discriminant validity was introduced by Campbell and Fiske (1959) within their discussion on evaluating validity. They stressed the importance of using both discriminant and convergent validation techniques when assessing new measures.
   A different view of construct validity is the idea developed by Lee Cronbach and Paul Meehl in 1955 of the *Nomological Network*. The "nomological" is derived from Greek and means "lawful". According to this idea, in order to test construct validity, a nomological network needs to be developed. This network would include the theoretical framework for what is going to be measured, an empirical framework for how it will be measured, and specification of the linkages among and between these two frameworks, as represented in the following figure (from Trochim, 2000):



The nomological network is founded on a number of principles that guide the researcher when trying to establish construct validity. They are:

- Scientifically, to make clear what something is or means, so that laws can be set forth in which that something occurs.
- The laws in a nomological network may relate:
    - observable properties or quantities to each other
    - different theoretical constructs to each other
    - theoretical constructs to observables
- At least some of the laws in the network must involve observables.
- "Learning more about" a theoretical construct is a matter of elaborating the nomological network in which it occurs or of increasing the definiteness of its components.
- The basic rule for adding a new construct or relation to a theory is that it must generate laws (nomologicals) confirmed by observation or reduce the number of nomologicals required to predict some observables.
- Operations which are qualitatively different "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable.

Cronbach and Meehl tried to link the conceptual/theoretical realm with the observable one, because this is the central concern of construct validity. While the nomological network idea may work as a philosophical foundation for construct validity, it does not provide a practical and usable methodology for actually assessing construct validity.

The next phase in the evolution of the idea of construct validity – the development of the *multitrait-multimethod* matrix – moved us a bit further toward a methodological approach to construct validity.

Different statistical approaches and strategies can be applied in order to assess validity (Maggino, 2007).

3. Technical issues of the measurement process

Quantitative measurement of subjective components needs to be founded on strong theoretical and methodological principles. These principles, grounded in psychometrics standards, states that the measurement of subjective characteristics requires a model in order to obtain interpretable and analysable information. The model, allowing *observation* (the collected information) to be transformed into *datum* (analysable information)[1], is composed by two sub-models:

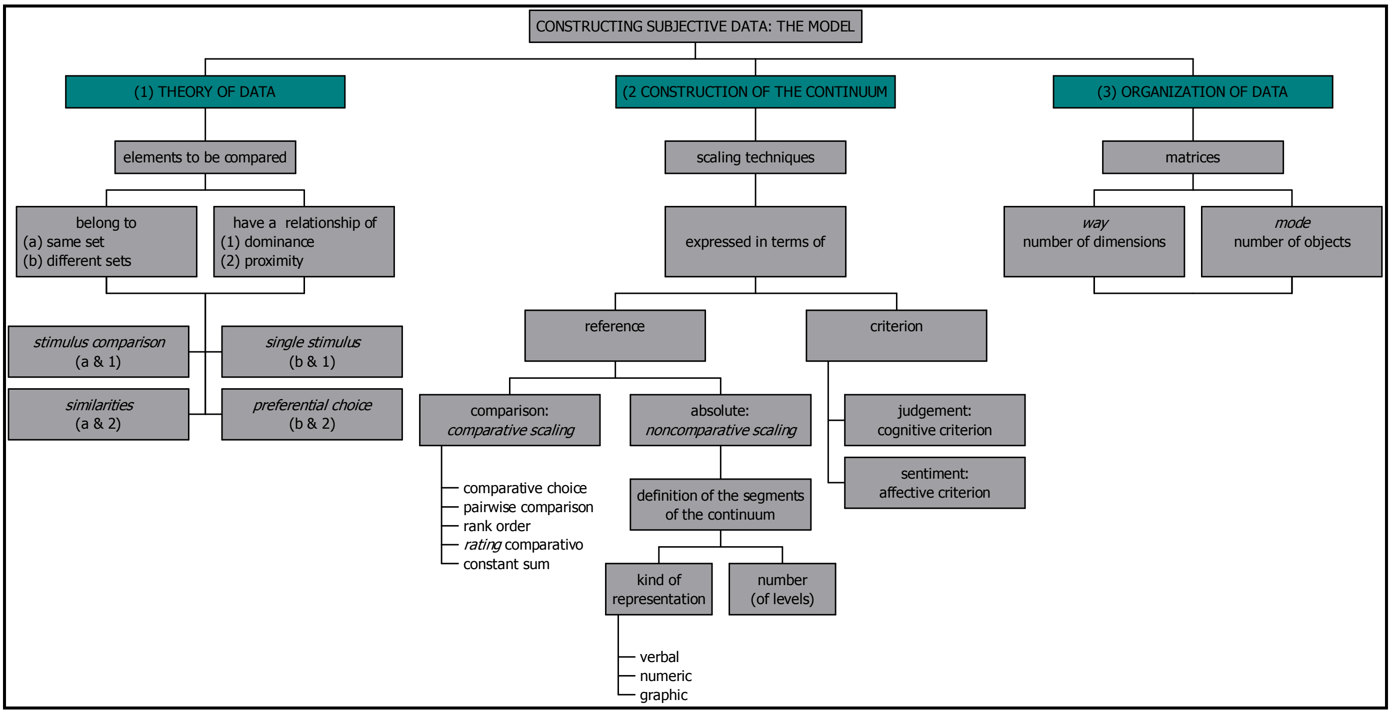
1. *model for constructing subjective data* (defining the *nature of data)*, which requires not only a theory of data and a method of data management, but also a procedure aimed at the construction of the continuum on which each unit has to be placed with reference to the observed characteristic,
2. *model for assigning data values* (defining the *rules for numeric assignment)*; this model allows a value to be assigned that makes the constructed data interpretable and that may be treated in operative terms (system of measurement).

This modelling is aimed at ensuring comparability within-individual and between-individuals.

3.1 Constructing subjective data

The definition of subjective data represents one of the more delicate stages of the measurement procedure and needs special care, in order to avoid excessive arbitrary elements and inaccuracy. The definition of subjective data requires a model defining:

1. nature of data, with reference to an interpretative theory (**theory of data**) (Coombs, 1950, 1953, 1964; Flament, 1976; Jacoby, 1991; McIver & Carmines, 1979),
2. procedure aimed at the construction of the continuum on which each individual case can be placed (**scaling techniques**) with reference to the observed characteristic (Amaturo, 1989; Andrews & Withey, 1976; Lodge, 1981; Marradi, 1980; Stevens, 1951, 1957; Weller & Romney, 1990),
3. data organization (**data matrix**) (Delli Zotti, 1995; Jacoby, 1991).



[1] Data consist of portions of information extracted according to a reference model; in this sense, data represent a researcher's construction and interpretation. As well-known, "data" represents the plural form of the Latin neuter term "datum" which represents a form (participle) of the Latin verb "dare" (do, das, dedi, datum, dare). The meaning of this verb is to assign, to fix, to establish, to frame, to place.

3.1.1 Nature of data

In order to convert empirical observation into understandable, interpretable and analysable information (datum), a theory is needed allowing nature of information to be clarified. The reference theory for subjective data definition is that defined by Coombs (1950, 1953, 1964; Flament, 1976; McIver & Carmines, 1979), based upon geometrical interpretation.
All empirical observations can be represented like comparisons, more or less explicitly performed,[2] between at least two entities, which can be defined as points into a particular space. The relative positions of the two points depend on the interpretation given to the comparison between the two entities.[3]
For each observation, the datum represents the portion of observation summarizing the comparison between entities. In other words, the datum can be defined in terms of geometrical relationship between two entities.
The geometrical representation of data composes a model.
Coombs (1953, 1964; Flament, 1976; Jacoby, 1991; McIver & Carmines, 1979) developed a theory based completely on the geometrical interpretation of data. According to the theory, two entities in a single datum can vary with reference to two criteria:

a. the set to which the entities belong to. The entities can belong to the same set (e.g., two individual who take the same test) or to two different sets (e.g., a stimulus and a response);
   - two *__different sets__* (e.g. a student and a test);
   - the *__same set__* (e.g. student *A* and student *B*).

   Determining whether entities belong to the same or different sets is possible by considering the substantive nature of the entities.[4]
b. the relation in which the entities are involved that can be:
   - *__dominance relation__*: an individual answers a question by reporting a level exceeding a defined measure;
   - *__proximity relation__*: two individuals share an event or two objects match or coincide with each other at different levels.

   Even if the difference between the two relations can be easily detected from the nature of empirical observation, the distinction rests with the analyst's interpretation of the observations.[5]

The two perspectives can be easily transformed in geometric representations: entities included in a single observation are always described as two points located within a space.[6] If two elements are drawn from
- different sets, the space is often called *joint space* (because it contains two distinct sets of points,
- a single set, the space is called *object space*.[7]

If the objects involved in the observation are related by a
- *dominance* relation, this is reflected by the ordering of the points in the space (if one entity dominates another, its point is placed at a higher position along the dimension)
- *proximity* relation, the objects are seen in terms of interpoint distance (more proximal two objects, smaller distance between them, and vice versa).

Combination between the two criteria (set and relation) produces four different kinds of data: all empirical observations, regardless of their substantive nature, can be classified into one of the four types (Flament, 1976; Jacoby, 1991):

		Pairs of points in observation	
		same set	different set
Relation between points in pair	dominance	Stimulus comparison **a**	Single stimulus **b**
	proximity	Similarities **c**	Preferential choice **d**

a. *__Stimulus comparison__*: observations are pairs of objects drawn from the same set with a dominance relation between them. This occurs when similar objects are compared to each other based on some

[2] The assertion «the apple is red» represents a comparison between the «apple» and a series of colours.
[3] With reference to the assertion «the apple is red», the two corresponding points – «apple» and «red» – are relatively close in the defined space.
[4] In some situation two entities, apparently belong to the same set, can be considered as if they belong to two different sets (e.g. the distinction between individuals that choose and individuals that are chosen.
[5] Let see an example. A subject completes some assigned tasks. Two interpretations are possible (Jacoby, 1991):
a. dominance ⟶ the subject's skill level exceeds that necessary for the task,
b. proximity ⟶ the subject's skill coincides with that necessary to complete the task.
[6] It should be taken into account that even if the space can be also multidimensional, here we will more simply refer to an uni-dimensional space.
[7] This space can be also called in other ways, according to the dealt object (*subject*, *stimulus*, etc.)

common property, describable by a dimension along which the observations can be modelled as an ordering of points. Some examples are:

b. Single stimulus: observations are pairs of objects drawn from different sets and with a dominance relation between them. The two point sets are (i) the objects being measured, and (ii) the calibration units on the measurement instrument. If object *A* shows a score *y* on scale *x* then *A* dominates all the scores up to *y* and fails in dominating scores greater than *y*.[8] Regardless of the meaning of the observations, the geometric model implies an order relation between pairs of points with reference to the underlying dimension. Virtually all physical measurement falls within this data category.
c. *Similarities*: observations are pairs of objects drawn from the same set and with proximity between them. The concept of similarity implies that two stimuli are judged to be more or less similar (proximity between them may increase or decrease). Two examples may be (i) the correlation between two variables and (ii) the degree to which two objects are confused through the judgments of a group of individuals. The information does not say anything about the ordering of the points within the space.
d. *Preferential choice*: observations are pairs of objects drawn from different sets and with a proximity relation between them. The most obvious example is represented by the case in which a given subject likes/prefers a particular stimulus: the more he/she likes, the greater the proximity between the subject and the stimulus. Another case is that in which stimuli are rated according to the degree that they exhibit certain characteristics. The more a stimulus possesses a characteristic, the greater the proximity between that stimulus and the characteristic. Geometrically, the proximities are represented as distances between points, within a joint space. Increasing proximity between a subject and a stimulus corresponds to decreasing distance between the subject-point and the stimulus-point. The information contained in any preferential choice data supplies no information about the relative ordering of the subject and the stimulus within the space.

3.1.2 Scaling techniques

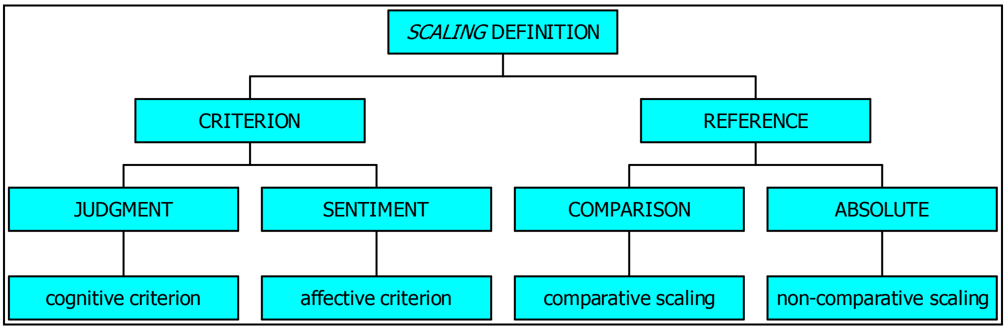
In constructing the subjective datum a problem is placed in order to define and, to some extent, to create and generating the continuum along which subjects can be located with reference to the measured characteristic. Identifying the continuum needs to take into account that till this point its definition is only theoretical.
This procedure, recalling what was defined at data theory level, is named *scaling*, which can be defined according to different approaches (Marradi, 1980). Actually, every approach assumes a continuum, which can be

- broken up into discrete ordinal categories (*discrete scaling*); in this case the continuum is defined *underlying continuum*. An assumption should be made about the extension of each category/segment on the continuum. Generally, it is “easier” to assume equal extensions (usually, and erroneously, this situation is defined “equally distanced categories”), later, we will see the procedures and the problems aimed at defining
- metrically defined (*continuous scaling*): all the positions identified on the continuum are related to each other through metrical properties; this kind of *scaling* emerges when between two points, much as they are close, it is always possible to identify another.[9] Since, this identification would require, in practice, a infinitely precise measurement, perfectly continuous scaling is an abstraction that is difficult to be observed.

As far as the presented arguments has shown, we can assert that the difference between the two approaches in identifying the continuum is in the knowledge/capacity to estimate/hypothesize the number and the extension of each segment (Marradi, 1980). Any *scaling* procedure requires the following aspects to be defined:

[8] Superficially, *single stimulus* and *stimulus comparison* may erroneously appear to be identical.
(i) *single stimulus*: one type of object is compared to some other (fundamentally different) kind of object. A stimulus is evaluated by a response (an object *A* possesses *x* units of one characteristic);
(ii) *stimulus comparison*: the comparison is made between two similar objects, drawn from the same set (a car has a better performance in terms of covered distance and fuel consumption (the dimension is represented by the rate value).
The outcome is the same in each case. For both types of data, the information implies the ordering of the points along the dimension. The difference between the two types concerns the information used to construct the geometric representation.

[9] The *magnitude scaling* (Lodge, 1981) represents one of the approaches allowing metrical scaling to be constructed. It is based upon the psychophysics theory (Stevens, 1951, 1957). Multidimensional correspondence analysis could represent another approach (analytical in its features) allowing metrical continuum to be identified (Amaturo, 1989; Weller & Romney, 1990).



3.1.2.1 The criterion

The reference refers to the kind of evaluation (**criterion**) urged at subjective level and connected to the studied characteristics. Two criteria can be distinguished:

- <u>*cognitive criterion*</u>: in this case a judgment or a knowledge is urged; the goal is that to set the relationship between *perceived* intensity and *actual* intensity of the characteristic. This means that the level of correctness and accuracy of the urged reaction can be checked. Evaluating similarities represents a typical example of cognitive criterion[10]*;*
- <u>*affective criterion*</u>: in this case a sentiment, a sensation, a preference, an interest, a sympathy, are urged; the goal is to determine the relationships between the measured characteristic and the individual; this kind of reference does not allow the level of correctness and accuracy to be assessed, making model development more complex.

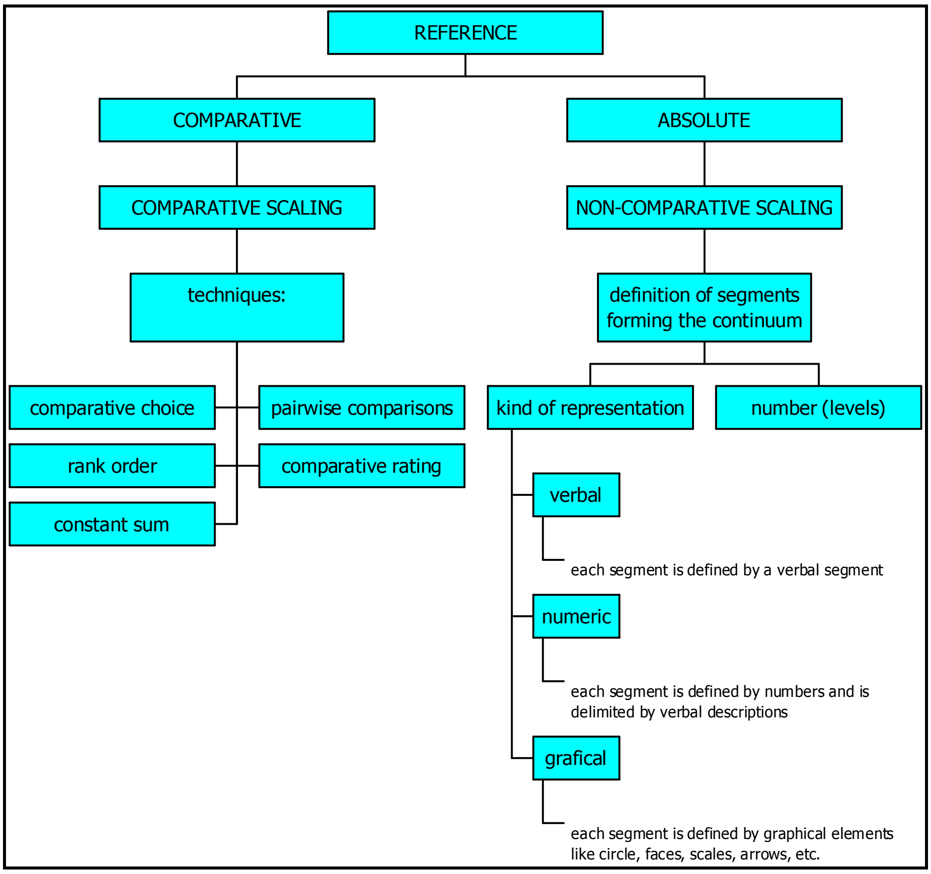
3.1.2.2 The reference

The submitted reference can be **comparative** or **absolute**.
From data collection's point of view (questionnaire administration), absolute references are preferable since they can be rapid submitted and produce more interpretable data to be obtained.
From data quality's point of view, comparative references allow more accurately evaluations to be obtained.[11]
According to the two references, two different scaling techniques can be arranged.

[10] Similarity criterion can include different concepts, *similarity/dissimilarity*, *relativity/generality*, *dependency/independency*, *association/separation*, *substitutability/irreplaceability*, *confusion/distinction*, and so on. Whatever concept is adopted, it is important to check

1. the respondent's possibility and capacity to use the concept in order to express consistent and comparable answers;
2. the certainty and truthfulness of the adopted concept with reference to the selected stimuli, in order to avoid difficulties in interpreting results.

[11] Concerning this, however it should be taken into account that the absolute reference is actually accomplished through a comparative process, even if not clearly required. In fact, in these cases, individuals are inclined to give "absolute" answers on the base of comparisons (concerning past experiences and activities, other persons, and so on).



Comparative scaling

Comparative scaling turns out to be very versatile. The subject is asked to compare two or more stimuli
- in relative terms → the subject reports his/her position concerning equality or difference between stimuli
- in ordinal terms → the subject reports the rank of each submitted stimulus.

The criterion for comparison can be cognitive or affective.
Main advantages of comparative scaling are:
- possibility to record also small differences between stimulus (since no tied rank is allowed);
- all the respondents use the same simple and easily applicable comparison procedure;
- no particular assumption is needed;
- reduction of halo effect among submitted stimuli.

The main disadvantage is represented by the difficulties in interpreting the ordinal nature of obtained data.
Practically, comparative scaling is accomplished through different techniques.

- ***Comparative choice***. This approach requires a series of stimuli (adjectives or sentences, etc.) to be defined and submitted. The respondent has to choose among them – through the given criterion – those that better describe a certain situation, feeling, figure, personality, and so on. *Adjective Check List* represents a typical comparative choice technique. The disadvantage of this technique is represented by the reduced possibility to analyse the obtained data (dichotomous in their nature: "chosen" / "not chosen").
- **Pairwise comparison.**[12] According to the given criterion (cognitive – similarity between stimuli – or affective – preference between stimuli), respondent is asked to choose between two stimuli or between three stimuli (triads). In both cases, all the possible objects' combinations have to be defined.[13] In some cases, respondents are allowed to express no similarity/preference (neutral response).
  The number of stimuli to be compared can represent a limit in using this technique. In fact, in presence of many stimuli, respondent's task becomes very complicated (with *k* stimuli, the number of pairwise comparisons is $k(k-1)/2$).

[12] Pairwise comparison can be applied also to objective information as it happens in recording the amount of communication between individuals (e.g. phone calls), groups, cities (e.g. train routes/trips). This kind of data can be used in order to map the communication and information flow by applying the MultiDimensional Scaling analysis.
[13] The similarity/preference could be expressed also in monetary or weight terms.

In certain cases, respondent can be asked to form group of objects (*clustering* technique) representing exhaustive and exclusive categories: objects belonging to the same categories must be very similar while those belonging to different categories must be dissimilar.

In order to apply pairwise comparison technique[14] consistency of the judgement has to be assesses by calculating the *consistency ratio* that expresses the internal consistency of the judgments that have been entered. To be called *consistent*, the rank should be transitive.[15] This allows pairwise comparison data to be turned into ranks.

MaxDiff represents a variation of this approach (Sawtooth, 2004) invented by Jordan Louviere in 1987 (1988). With MaxDiff, respondents are shown a set of the possible items and are asked to indicate the best and worst items (or most and least important, or most and least appealing, etc.). MaxDiff assumes that it is much more comfortable judge items at extremes than discriminate among items of middling importance or preference. Actually, respondents evaluate all possible pairs of items within the displayed set and choose the pair that reflects the maximum difference in preference or importance.

- **Rank order**. Respondent is asked to rank the stimuli with reference to the adopted criterion (cognitive or affective). This technique is simpler and thriftier (with *n* stimuli, rank order technique requires $n-1$ decisions while pairwise comparison technique requires $n(n-1)/2$ comparisons) than the previous one, even because the consistency assessment is not required.
  In order to apply correctly the technique, the respondent should be enabled to carry out the task, that is
  - the respondent should know all the objects to rank,
  - the number of objects should not be high.
- **Comparative rating**. Respondent is asked to report for each object a score related to the adopted criterion. The score can be reported in proportional terms or in percent terms.
- **Costant sum**. Respondent is asked to arrange a certain amount (money, time, etc.) among the stimuli with reference to the adopted criterion. Constant sum allows clear and interpretable data to be obtained but requires great accuracy in carrying out the task (the final sum of the assigned scores should correspond to the initial defined amount). Moreover, the number of objects should not be high.

Non-comparative scaling

By adopting an absolute reference, each respondent reacts (by using a previously defined scheme) to each stimulus independently by the others. This defines non-comparative scaling techniques.

Generally, non-comparative scaling techniques are very simply to be constructed and applied and can be classified with reference to:

a. **kind of representation** (verbal, numeric or graphical) of the continuum (Aureli & Koch-Weser, 1977). Choosing the kind of representation should take into account the adopted survey method (face-to-face interviews, telephonic interviews, presence of interviewers, etc.)
b. **anchoring**
c. **number of segments.**[16]

Subsequently, in order to make data useable for statistical analysis, a value should be assigned to each segment. The set of values defines the system of measurement.

(a) Kind of representation

Verbal representation. In this case, each identified segment is defined by a verbal label, showing the segment meaning. One of the examples of this kind of representation is represented by the widely known Likert scale (form its author's name, Rensis Likert, 1932). In this case, the continuum, representing the "agreement" (cognitive reference), can be subdivided as following:

strongly agree	slightly agree	slightly disagree	strongly disagree

or

strongly disagree	disagree	neither agree or disagree	agree	strongly agree

As we will argue later, some discussion exists concerning the presence of the central-neutral point. Generally, this scaling technique, is used in measuring attitudes, values, or opinions, when the respondent is asked to report their level of agreement to a statement.

The continuum can be verbally represented also with reference to the concept of "frequency" (*always – often*

[14] One of the most useful applications of this technique is the *Analytic Hierarchy Processes*, one of the approaches allowing weighting system to be defined.

[15] E.g., if object *A* is preferred to object *B*, and object *B* is preferred to object *C*, then object *A* is preferred to object *C*.

[16] The effect of representation, polarity and number of levels on data quality is often investigated (Dawes, 2008; Maggino, 2003; Lozano et al., 2008).

– *sometimes* – *rarely* – *never* or *never* – *shortly* – *for sometime* – *for long time*).
The defined scale and its segments need to be assessed with reference to:

a. segments' ordering,
b. segments' polarity (from agree to disagree),
c. segments' balance and symmetry.

The assessment is conditioned by and depends on the linguistic and cultural contexts.

Numeric representation. In this case, each identified segment is defined by a number. Using numbers could help in

- identifying the continuum and the segments' regularity
- avoiding the typical misinterpretations of verbal representations.

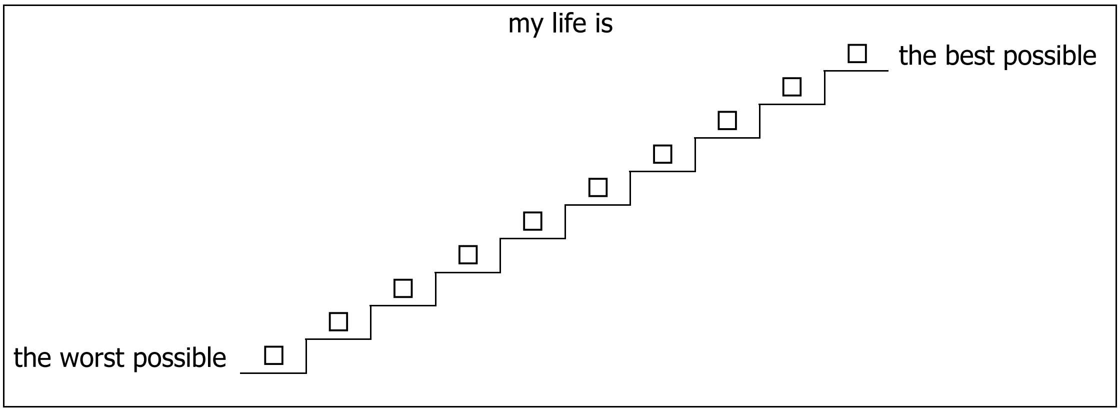
For example, subject may be asked to identify his/her state according to the given criterion by pointing out the corresponding value from "0" (the worst state) to "10" (the best state).
Particular care should be paid to the numerically represented continuum with reference to

- the polarity (for the lowest value up or from the highest value down)
- the identification of the polarity (i.e. anchoring)
- the quantitative meaning of numbers.

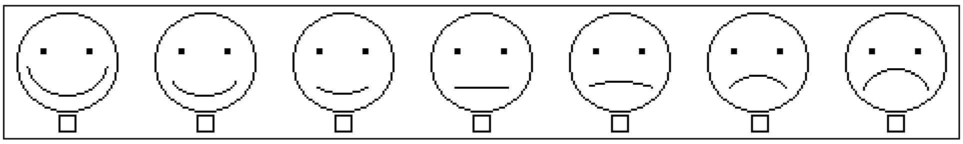
Sometimes, the numeric scaling is defined *rating*.

Graphic representation. As we have see, different problems may be originated by both verbal (e.g. semantic interpretation) and numeric (e.g. anchoring) representations. These problems increase in passing from one language to another (translating problems) and from one population to another (cultural problems). In many case, these problems can be unravelled by means of graphic representations, allowing different flexible solutions. Generally, graphic representation allows continuum to be more clearly communicated and subjects' task to be facilitated. Some examples, frequently applied in subjective well-being surveys, can be useful.

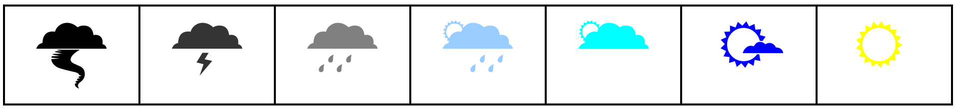
- *Ladder scale*: the segments describing the continuum are represented by stairs (usually with 9 or 11 rungs):

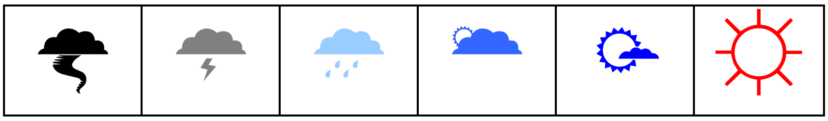


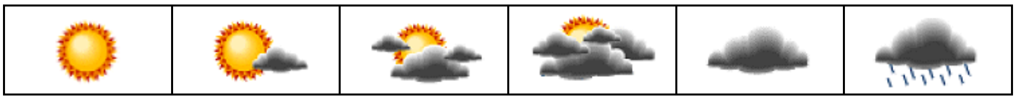
- *Faces scale*: the segments describing the continuum (referring to affective criterion) are classically represented by 7 stylised faces, which differ in mouth inclination, expressing different emotional states:



- *Weather scale*: the segments describing the continuum (referring to affective criterion) are represented by different weather phenomena recalling different emotional states:

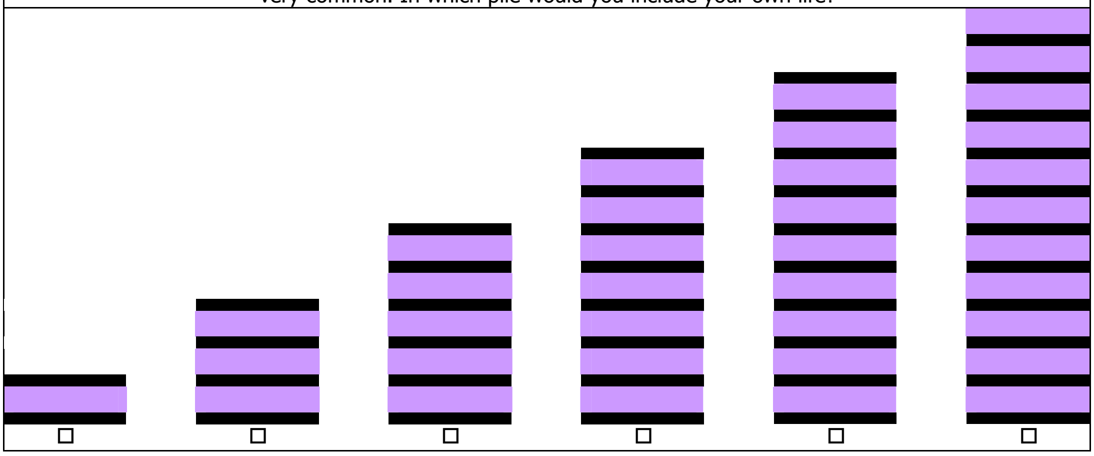






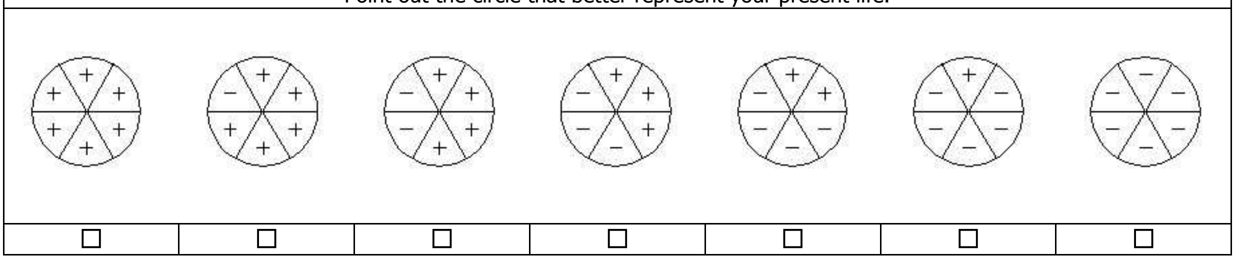
- *Pile scale*: the segments describing the continuum are represented by piles of different heights:

Image that each of the following piles represents a group of persons having the same life style. The first pile refers to persons whose life style is uncommon; on the contrary, the last pile refers to persons whose life style is very common. In which pile would you include your own life?



- *Circle scale:* the segments describing the continuum are represented by circles divided into slices. The circles gradually differ in the number of "+" and "-" occupying each slice (from the circle full of "+" to the circle full of "-"). The final representation produces symmetrical and bipolar segments.

Each of the following circles represents different life ambits. Each ambit can be positive ("+") or negative ("-"). Point out the circle that better represent your present life.



Particular care should be paid to the numerically represented continuum with reference to
- the polarity (for the lowest value up or from the highest value down)
- the identification of the polarity (i.e. anchoring)
- the meaning of the used graphic symbols
- the orientation (e.g. horizontal or vertical)

also because of different cultural effects.

A particular problem in graphic representation may be represented by the segments' alignment. The segments can be
- *directly connected*, by preserving the continuum idea also graphically:

agreement → | ---- ---- ---- ---- ---- ---- | ← disagreement

agreement → | ---- | ---- | ---- | ---- | ---- | ---- | ← disagreement

- *separated*, allowing respondent to point out more clearly the answer:

agreement → ☐ ☐ ☐ ☐ ☐ ☐ ← disagreement

agreement → ☐ ☐ ☐ ☐ ☐ ☐ ← disagreement

(b) Anchoring

In both numeric and graphic representations, both ends of the continuum should be (verbally) described. As we will see, the polarity is related to the anchoring issue. With reference to this, we can distinguish between:

- *scaling explicitly anchored*: the extremities are described by words/sentences (anchoring agents) that fix the meaning of the represented scale, like in the following example:

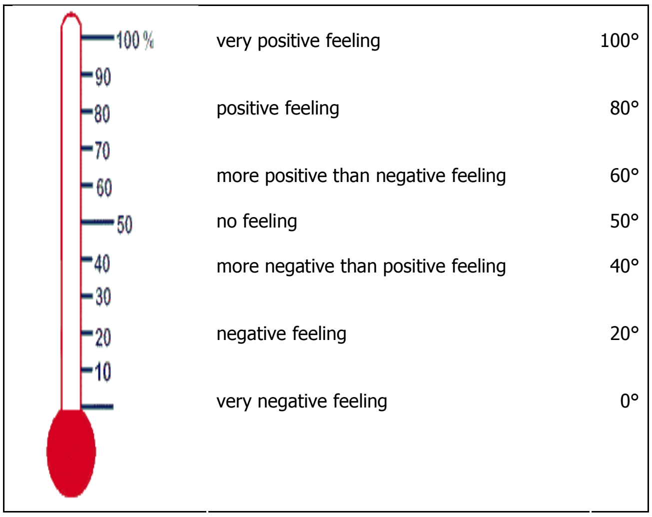
  agreement → ☐ ☐ ☐ ☐ ☐ ☐ ← disagreement

  Also the differential semantic approach (Maggino & Mola, 2007) typically uses an anchored scale:

Thinking about your city, point out the position that is closer to the adjective that better describes your ideal city.								
Silent	☐	☐	☐	☐	☐	☐	☐	Noisy

  In many cases, as it happens in differential semantic scaling, anchoring definition does not automatically allow the right polarity to be identified, especially because polarity is not only a technical but also semantic and cultural issue.

  Anchoring may involve not only the extremities of the scale but also other segments, like it happens in the following example representing the *feeling thermometer* (affective criterion). Respondent should point out on the thermometer, the score related to his/her feelings towards a given object (individuals, concepts, ideas, cities, institutions, etc.):



- *self-anchored scaling*: the extremities of the continuum are not clearly indicated but are determined by the respondent. This approach, even if showing some advantages, may produce scores not comparable between subjects.

(c) Number of segments

Defining the optimal number of segments in which the continuum should be subdivided is not a simple task. From one side, we know that the number of segments is positively and monotonically related to the level of reliable. In other words, a high number of segments helps in better discriminating among respondents. On the other side, the level of reliability does not increase when the number of segments is particularly high. In fact, respondent may find difficult to choose the segment that better expresses his/her position. The consequence is polarization of the answers towards some particular positions. This is particularly true with numeric representation: subjects could be induced to use only some scores (5, 10, 20, 50, etc.) nullifying the idea of precise measurement.

Other important issues that should take into account in defining the number of segments are:

- correct balance between positive and negative definitions – especially with verbal representation;
- odd or even number: odd numbers allows an intermediate segment to be introduced with the meaning of "neutrality". This median segment could make subjects' responses more comfortable. However, the

presence of the median segment could encourage respondent in avoiding a clear position.

At the end of this paragraph, it should be always taken into account that scaling technique's success depends also on

- correctness of instructions provided to respondents,
- respondents' honest and truthful attitude to answer (no *response set*),
- respondents' past experiences.

3.1.3 Data organization: matrices

Modelling data process requires also the definition of data organization (Delli Zotti, 1985). Carrol, Arabie e Young[17] (Jacoby, 1991) have defined a classification system of the different types of matrices, characterized by two features:

- **way**, referring to the *dimensions* of the matrix (number of indices used in identifying the objects). Each *way* has its number of *levels*, corresponding to the number of entities in that object set. Consequently, the *ways* define the shape of the data matrix while the levels specify the size of the matrix. Any matrix should be at least *two-way* since each observation always involves a comparison between two objects;
- **mode**, referring to the *number of objects* represented by the *ways* of the matrix. *Modes* determine the interpretation of the objects. The number of modes depends on the type of objects. The number of *modes* cannot exceed the number of *ways*.

- **Two-way one-mode matrix**. Elements defining the rows correspond to elements defining the column: the matrix contains a single set of information (one mode). This is the typical case of the <u>square matrices</u>, like in the following examples:
    - <u>Correlations between indicators</u> (r_{ij} = correlation between indicators *i* and *j*):

two-way one-mode matrix:
indicators * indicators

indicators

indicators	1	2	...	j	...	k
1	r_{11}	r_{12}	...	r_{1j}	...	r_{1k}
2	r_{21}	r_{22}	...	r_{2j}	...	r_{2k}
...	...	...	...	...	...	...
i	r_{i1}	r_{i2}	...	r_{ij}	...	r_{ik}
...	...	...	...	...	...	...
k	r_{k1}	r_{k2}	...	r_{kj}	...	r_{kk}

    - <u>Similarities between cases</u> (s_{ij} = similarity between cases *i* and *j*):

two-way one-mode matrix:
case * case

cases

cases	1	2	...	j	...	n
1	s_{11}	s_{12}	...	s_{1j}	...	s_{1n}
2	s_{21}	s_{22}	...	s_{2j}	...	s_{2n}
...	...	...	...	...	...	...
i	s_{i1}	s_{i2}	...	s_{ij}	...	s_{in}
...	...	...	...	...	...	...
n	s_{n1}	s_{n2}	...	s_{nj}	...	s_{nn}

- **Two-way two-mode matrix**. It is the most common matrix: if we have *k* indicators on *n* cases, we will

[17] This theory is often known through the authors' initials (*CAY*).

have a matrix with two modes (cases and indicators) – two ways matrix. The first way has *k* levels, while the second way has *n* levels. In the following matrix, v_{ij} represents the value observed by case *i* for indicator *j*.

Two-way two-mode matrix: the typical data matrix

cases \ indicators	1	2	...	j	...	k
1	v_{11}	v_{12}	...	v_{1j}	...	v_{1k}
2	v_{21}	v_{22}	...	v_{2j}	...	v_{2k}
...	...	...	...	...	...	...
i	v_{i1}	v_{i2}	...	v_{ij}	...	v_{ik}
...	...	...	...	...	...	...
n	v_{n1}	v_{n2}	...	v_{nj}	...	v_{nk}

The following matrix is another example of a two-way two-mode matrix.

Two-way two-mode matrix

Time points \ indicators	1	2	...	j	...	k
1	v_{11}	v_{12}	...	v_{1j}	...	v_{1k}
2	v_{21}	v_{22}	...	v_{2j}	...	v_{2k}
...	...	...	...	...	...	...
i	v_{i1}	v_{i2}	...	v_{ij}	...	v_{ik}
...	...	...	...	...	...	...
t	v_{t1}	v_{t2}	...	v_{tj}	...	v_{tk}

- **Three-way three-mode matrix**. If we have *n* cases, *k* indicators, *m* time points, the matrix has
  - three modes: (1) cases (*n* levels), (2) indicators (*k* levels), and (3) time points (*m* levels)
  - three ways: the third way refers to the repeated observations.

Three-way three-mode matrix

cases	Time point 1 indicators 1	2	...	j	...	k
1	$v_{1.11}$	$v_{1.12}$	...	$v_{1.1j}$	...	$v_{1.1k}$
2	$v_{1.21}$	$v_{1.22}$	...	$v_{1.2j}$	...	$v_{1.2k}$
...	...	...	...	...	...	...
i	$v_{1.i1}$	$v_{1.i2}$	...	$v_{1.ij}$	...	$v_{1.ik}$
...	...	...	...	...	...	...
n	$v_{1.n1}$	$v_{1.n2}$	...	$v_{1.nj}$	...	$v_{1.nk}$

cases	Time point 2 indicators 1	2	...	j	...	k
1	$v_{2.11}$	$v_{2.12}$	...	$v_{2.1j}$	...	$v_{2.1k}$
2	$v_{2.21}$	$v_{2.22}$	...	$v_{2.2j}$	...	$v_{2.2k}$
...	...	...	...	...	...	...
i	$v_{2.i1}$	$v_{2.i2}$	...	$v_{2.ij}$	...	$v_{2.ik}$
...	...	...	...	...	...	...
n	$v_{2.n1}$	$v_{2.n2}$	...	$v_{2.nj}$	...	$v_{2.nk}$

[...]

cases	Time point *m* indicators 1	2	...	j	...	k
1	$v_{m.11}$	$v_{m.12}$	...	$v_{m.1j}$	...	$v_{m.1k}$
2	$v_{m.21}$	$v_{m.22}$	...	$v_{m.2j}$	...	$v_{m.2k}$
...	...	...	...	...	...	...
i	$v_{m.i1}$	$v_{m.i2}$	...	$v_{m.ij}$	...	$v_{m.ik}$
...	...	...	...	...	...	...
n	$v_{m.n1}$	$v_{m.n2}$	...	$v_{m.nj}$	...	$v_{m.nk}$

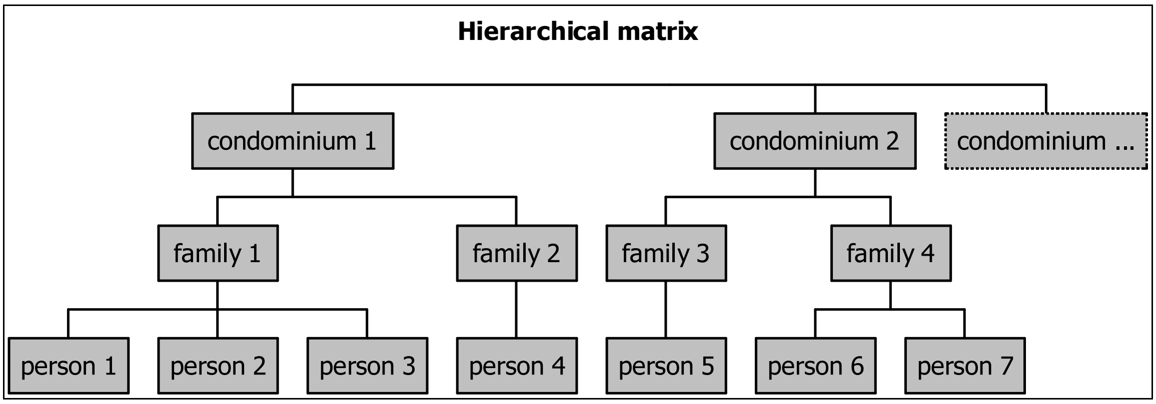
- **Three-way two-mode matrix**. The following table represents an example of this kind of matrix

		Time point 1 indicators						Time point 2 indicators							Time point m indicators					
		1	2	...	j	...	k	1	2	...	j	...	k		1	2	...	j	...	k
Three-way two-mode matrix	indicators 1	$r_{1.11}$	$r_{1.12}$	...	$r_{1.1j}$	...	$r_{1.1k}$	$r_{2.11}$	$r_{2.12}$	...	$r_{2.1j}$	...	$r_{2.1k}$		$r_{m.11}$	$r_{m.12}$	...	$r_{m.1j}$	...	$r_{m.1k}$
	2	$r_{1.21}$	$r_{1.22}$	...	$r_{1.2j}$	...	$r_{1.2k}$	$r_{2.21}$	$r_{2.22}$	...	$r_{2.2j}$	...	$r_{2.2k}$	[...]	$r_{m.21}$	$r_{m.22}$	...	$r_{m.2j}$	...	$r_{m.2k}$
	...	...	...	...	...	...	...	...	...	...	...	...	...		...	...	...	...	...	...
	i	$r_{1.i1}$	$r_{1.i2}$	...	$r_{1.ij}$	...	$r_{1.ik}$	$r_{2.i1}$	$r_{2.i2}$	...	$r_{2.ij}$	...	$r_{2.ik}$		$r_{m.i1}$	$r_{m.i2}$	...	$r_{m.ij}$	...	$r_{m.ik}$
	...	...	...	...	...	...	...	...	...	...	...	...	...		...	...	...	...	...	...
	k	$r_{1.k1}$	$r_{1.k2}$	...	$r_{1.kj}$	...	$r_{1.kk}$	$r_{2.k1}$	$r_{2.k2}$	...	$r_{2.kj}$	...	$r_{2.kk}$		$r_{m.k1}$	$r_{m.k2}$	...	$r_{m.kj}$	...	$r_{m.kk}$

In the same way, a **four-way four-mode matrix** can be defined where we have *n* cases, *k* indicators, *q* areas, *m* time points, the matrix has

- four modes: (1) cases (*n* levels), (2) indicators (*k* levels), (3) time points (*m* levels), and (4) areas (*q* levels)
- four ways: the fourth way refers to the areas.

Data matrices can present also particular hierarchical structures. In these cases, same cases/units[18] are subordinated to others, according to a tree structure. A typical example of a **hierarchical matrix** is that showing three kinds of elements: cases: subjects, families, condominiums. In this structure, each individual is associated to one family (and its information) and each family (and its information) to one condominium (and its information):



Within a hierarchical structure, each unit category can be considered from the analytical point of view and gives rise to a different two-way two-mode matrix. By referring to the example, we could obtain three different matrices in which each column represents an indicator and each row represents respectively

- one condominium, carrying on information concerning all the families within it and all the individuals within each family,
- one family, carrying on information concerning the condominium to which it belongs and all the individuals within it,
- one individual, carrying on information concerning the condominium and the family to which he/she belongs.

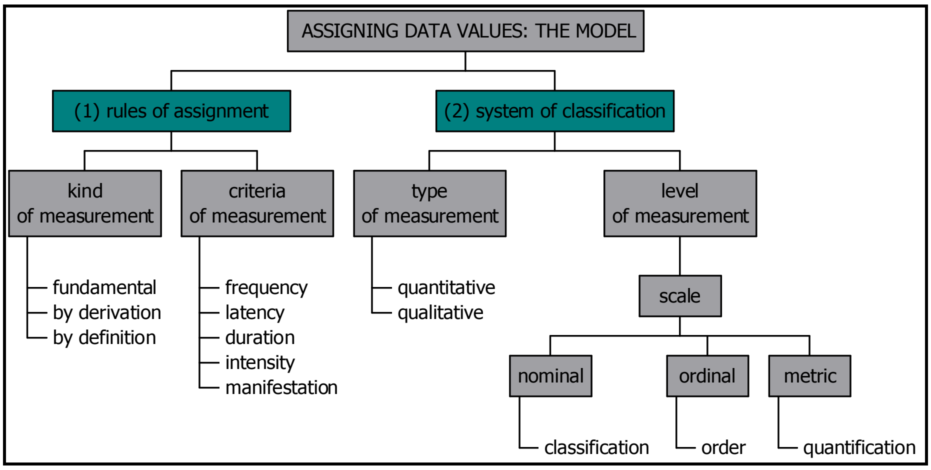
The adequacy of the matrix to any given situation depends entirely upon the analysist's interpretation of the observations. Moreover, the number and type of entities contained in the observations may or may not correspond to the shape and size of the data matrix. The previous example shows how the data are treated as an abstract model extracted from the observations. Per se, the characteristics of the data are entirely independent from the substantive properties of the observations themselves.
Further, the CAY data theory can be considered completely consistent with the Coombs fourfold theory

[18] *Unit* is an individual case, object of study. It can be represented by individual, family, firm, and so on. Each case can be observational unit – when object of observation - or analysis unit – when object of analysis.

(Jacoby, 1991).[19] For this reason, it could be useful to conceive of data by referring to both theories since each theory clarify different aspects of information.

3.2 Assigning data values

Following the definition of the model for constructing the datum, the definition of a model is required in order to assign analysable values to observed data consistently to the identified continuum. This model needs to allow a value to be assigned that makes the constructed data interpretable and that can be treated in operative terms.
For this purpose, we need to define the rules clarifying the procedure of correspondence and of assignment of a symbol to each identified level. This requires the definition of a system of measurement that presents:
1. rules that allow numbers/symbols to be assigned in a standard and uniform procedure (**kind and criteria of measurement**) (Bruschi, 1999),
2. a "system of classification" that allows the status with reference to the measured characteristic to be assigned to each case (**type and level of measurement**) (Caracciolo in Siegel & Castellan, 1992; Stevens, 1946, 1951; Velleman & Wilkinson, 1993).



The whole group of values identified according to this model defines what is usually called a *scale*.[20]
The definition of the "system of measurement" represents one of the most debated points concerning the subjective measurement. In fact, except for those rare cases in which the system is self-evident and does not require a detailed formulation, the definition of the "system of measurement" is never simple, unambiguous, clear, or intuitive, and it usually raises problems of arbitrariness.

[19] *Single stimulus* and *preferential choice* data both produce matrices with at least two ways and two modes. The differences among them concern the comparative relation that holds across the modes: it is a dominance relation in the first case, and a proximity relation in the second case. *Stimulus comparison* and *similarity* data both generate two-way matrices with a single mode: again, it is a dominance relation in the first case, and a proximity relation in the second case. Of course, it is possible to have replicated observations for any of these kinds of data: in this case, the number of ways and modes in the data matrix will increase accordingly.

[20] Another strategy that allows to assign a value to each segment and that does not require the definition of an a priori system is the *optimal scaling* that proceeds through a particular analysis approach; the score to be assigned needs to meet two simultaneous conditions; it has:
- to fit the statistical model as well as possible and
- to preserve strictly the specified characteristics of the measurement.

The *optimal scaling* strategy presents the advantage to provide for the best series of numeric assignments by assessing the fitness between analytical model and the empirical observations. While the model of measurement has to be specified in advance, the researcher can vary the assumptions by observing the effects produces by the analytical approaches: if numerous analyses produce exactly the same results then the adopted assumptions are those referring to the analyses that have produced equivalent results. In this sense, the *optimal scaling* strategy explicitly embodies the concept of measurement as verifying theories process.

4. *Synthesizing subjective indicators: the scaling models*

As we have seen, the model generating subjective indicators is the factor model. Applying factor analysis allows indicators' communalities to be checked: high values in communalities allow synthetic indicators to be computed. However, in order to create synthetic indicators, particular models have to be applied that enable:

- the conceptual model to be checked,
- the unity of the concept of interest to be re-established meaningfully,
- the multiple measures to be synthesized and the obtained synthetic value assigned to each individual,
- the continuum on which each individual can be placed in a meaningful, interpretable and manageable way to be identified.

The reflective measurement model has its roots in classical test theory and psychometrics (Lord and Novick, 1968; Nunnally, 1978). The methodology on scale development (Spector, 1990; DeVellis, 1991; Netemeyer et al., 2003) is based on reflective measures, examined in order to verify the main scale properties, including dimensionality, internal consistency and convergent/discriminant validity. In this perspective, a wide range of techniques of scale construction and measurement assessment (named *scaling models*) are applied which refer more or less to factor analysis. Scaling models can be defined as a design in order to consistently develop a new measure (Nunnally, 1978).

The scaling models can be distinguished with regard to different elements (McIver & Carmines, 1979, Maggino, 2007):

- **Dimensionality**. It is related to the complex nature of the defined latent variable; each dimension is related to different aspects of the defined variable. The identification of a certain dimensionality requires the adoption of a scaling model (McIver & Carmines, 1979; Netemeyer et al., 2003). The concept of "dimensionality" is quite complex, because its meaning is mainly and essentially theoretical. Two different dimensionalities can be distinguished:
  a. <u>uni-dimensionality</u>: in this case, the definition of the considered variable assumes a unique, fundamental underlying dimension;
  b. <u>multidimensionality</u>: in this case, the definition of the considered variable assumes several underlying aspects (dimensions).[1]

  The correspondence between the defined dimensionality and the selected elementary indicators has to be demonstrated empirically by testing the selected scaling model.
- **Nature of data**. As previously mentioned, the nature of data is not predetermined but depends on the researcher's interpretation, expressed in terms of appropriateness and consistency. Different interpretations lead to different scaling procedures. The different scaling procedures can be distinguished according to the classical classification of subjective data, theorized by Coombs (Coombs, 1950, 1953, 1964; Flament, 1976; Jacoby, 1991; McIver & Carmines, 1979):
  - <u>*Single stimulus*</u>. Many scaling models were conceived for this kind of data; they are very often applied, such as the *additive model* and the *cumulative models* (deterministic and probabilistic) (Flament, 1976; McIver & Carmines, 1979; Torgerson, 1958).
  - <u>*Stimulus comparison*</u>. Reference scaling models for this kind of data are the Thurstone model (Arcuri & Flores D'Arcais, 1974; McIver & Carmines, 1979; Thurstone, 1927, 1959) and the Q methodology (McKeown & Thomas, 1988).
  - <u>*Similarities*</u>. The reference scaling model for this kind of data is the *multidimensional scaling* (Cox & Cox, 1994; Kruskal & Wish, 1978; Torgerson, 1958).
  - <u>*Preferential choice*</u>. One of the reference scaling models is the *unfolding model* (McIver & Carmines, 1979).
- **Scaling technique**, comparative or non-comparative (Maggino, 2007).
- **Criterion for testing the model**. It is aimed at checking model data fitting. The procedure rationale is common to all the models but criteria show different characteristics, according to the chosen model (Maggino, 2007).
- **Standard of measurement**, concerning the treatment of the multiple measures and the assignment of the synthetic value (the final score can be assigned to individuals or to stimulus), according to the

[1] The notion of dimensionality is present in social sciences but also in others; concerning this, we can refer to unidimensional attributes as length and weight and multidimensional attributes like color and space.

following pattern:

Standard of measurement		Multiple measures	With regard to the variable the objective of the measurement is to classify	Final score assigned to
The multiple measures allow to measure in more accurately	individual	**Stimulus (item)**	the individuals	**Individual**
	indicator	**Individual**	the elementary indicators	**Stimulus (item)**

The following example allows us to understand the role, the weight and the meaning that each individual answer can assume according to the standard of measurement.

E.g. in a study on social prejudice, one variable is the "perception of the social distance from a defined social group "; in this case, the multiple measures can be represented by different items constituted by sentences concerning particular hypothetical behaviors towards the members of that social group ("I don't want anything to do with him/her", "I would accept sitting besides him/her on the bus", "I would accept him/her as a colleague", "I would invite him/her home", "I would accept him/her as a friend", "I would accept him/her as relative in-law"[2]); each individual expresses his/her agreement ("yes") or not ("no") regarding each behavior.

If the goal is to measure the individual level of the perceived social distance, the multiple measures should be represented by the whole set of items (that is, the whole group of answers given by a certain individual case to the whole set of items can be synthesized and allows the individual case to be placed on the "perceived social distance" continuum).

If the goal is to measure the level of social distance that each item is able to detect, the multiple measures should be represented by the whole group of individuals (that is, the whole group of answers obtained for a certain item from the whole group of individual cases can be synthesized and allows the item to be placed on the "perceived social distance" continuum).

- **Contribution of each multiple measure to the measurement**: the contribution can be *uniform* (that is, all the multiple measures contribute through the same evidence) or *differential* (that is, the multiple measures contribute through different evidence); in this perspective, a particular item characteristic can be considered, the *trace line*, that defines the relationship between the identified continuum and the frequency observed for each value of that continuum. This frequency can be interpreted in terms of "probability to obtain each value" (McIver & Carmines, 1979). In particular, two frequency distributions can be associated to each item, corresponding to two different probabilities respectively:
  - *alpha*, probability relating to the expected value ("correct answer " or "agreement with the submitted sentence" o "answer that is in the direction of the measured variable");
  - *beta*, probability relating to the not-expected value ("incorrect answer" or "disagreement with the submitted sentence" or "answer that is in the opposite direction to the measured variable").

The following table (Maggino, 2007) summarizes the characteristics of the well-known scaling models:

[2] This example refers to the *Bogardus Social Distance Scale*, a psychometric instrument created by Emory S. Bogardus to empirically measure people's willingness to participate in social contacts of varying degrees of closeness with members of diverse social groups. The Bogardus Social Distance Scale is based upon a cumulative scaling model, because agreement with any item implies agreement with all the preceding items (Maggino, 2007).

				Scaling models' characteristics				
				Dimensionality	**Nature of data**	**Scaling technique**	**Criterion for testing the model**	**Standard of measurement: final (synthetic) score assigned to**
Scaling models	**Additive**	Uni-dimensional		Uni	Single-stimulus	Not-comparative	Internal consistency	Cases
		Multidimensional		Multi	Single-stimulus	Not-comparative	Dimensionality of the items	Cases
	Cumulative	Thurstone model (differential scale)		Uni	Stimulus comparison	Comparative (pair comparison or rank-order)	Metrics between items	Items
		Q methodology		Uni	Stimulus comparison	Comparative (rank-order or comparative rating)		Items
		Deterministic	Guttman	Uni	Single-stimulus	Not-comparative	Scalogram analysis: reproducibility, scalability and ability to predict	Cases and items
			Multidimensional Scalogram Analysis (MSA)	Bi			Regionality and contiguity	Cases and items
			Partial Ordered Scalogram Analysis (POSA)	Bi			Correct representation	Cases and items
		Probabilistic	Monotone (one or more parameters)		Single-stimulus	Not-comparative	• parameters estimation (maximum likelihood) • goodness of fit (misfit and residuals analysis)	Cases and items (without condensation)
	Perceptual Mapping	Multidimensional scaling		Multi	Similarities	Comparative (pair comparison)	Goodness of fit of distances to proximities (stress, alienation)	Items
		Unfolding		Uni & Multi	Preferential choice	Comparative	Goodness of fit of distances to ordinal preferences	Cases and items
	Conjoint model			Multi	Preferential choice	Comparative (rank-order)	Goodness of fit of the model (part-worth) to the ranking	Items at individual level

4.1 Additive model

Additive model is based upon two important assumptions concerning the nature of indicators that are caused by only one latent variable and are linearly related to the latent variable (Spector, 1992).
In order to adopt this model, indicators should be selected according to their capacity to discriminate among cases from their position along the underlying continuum. Besides, the indicators should present homogeneous scaling techniques, in other words the same classifications should be applied to whole group of indicators.
In order to test the goodness of fit of the model, the main objective is to verify the presence of a common variance (Carmines & Zeller, 1992; Spector, 1992; Traub, 1994). This can be done through different approaches:

- *components based* approach: the indicators are subdivided into two components – which are or not perfectly equivalent (respectively parallel and non-parallel); the most common dividing technique is the split-half. The procedure aims at verifying the equivalence between components. The most used equivalence measures are the *Spearman-Brown* coefficient (parallel components) and the *Rulon* coefficient (non-parallel components),
- *indicators based* approach: each indicator is considered a component. In order to test the goodness of fit the internal consistency procedure is applied, evaluated through different coefficients (Cronbach's *alpha*, *KR-20*, *KR-21*, *alfa*, L_2).

<table>
<tr><th colspan="2" rowspan="2">APPROACH</th><th colspan="3">MODEL TESTING</th><th rowspan="2">PROBLEMS</th></tr>
<tr><th>KIND OF TESTING PROCEDURE</th><th>METHODS</th><th>TECHNIQUES AND INSTRUMENTS</th></tr>
<tr><td rowspan="2">Components</td><td>parallel</td><td rowspan="2">equivalence between the two components (split-half)</td><td rowspan="2">comparison between components</td><td rowspan="2">▪ Spearman-Brown coefficient
▪ correlation between components
▪ Rulon coefficient</td><td>identification of the parallel components</td></tr>
<tr><td>non-parallel</td><td>Identification of the components</td></tr>
<tr><td colspan="2">Internal consistency analysis</td><td>comparison between n components</td><td>comparison between
▪ indicators
▪ each indicator and the whole group</td><td>▪ correlation between indicators
▪ correlation indicator-total
▪ coefficients: alpha, KR-20, KR-21, L_1, L_2</td><td>identification of homogeneity of indicators</td></tr>
</table>

This model shares some characteristics with the factor model. In particular, the following assumptions:

- correlations between indicators are explained only by the presence of the latent variable,
- measurement errors are uncorrelated to each other, and are not correlated with latent variable,
- measurement errors are random.

The application of factor analysis in order to test the additive model allow the definition of parallel components to be avoided, in fact, *factor loadings* allow us to determine the contribution of each indicator to factor structure definition.
Additive and factor models are distinguished with reference to the definition of 'error': the latter defines only one random component, the former two components, which, however, are jointly estimated (*uniqueness*).

4.2 Cumulative models

In order to measure characteristics that are cumulative in their nature (e.g. capacities, perception of social distance, dispositions, difficulties, and so on), several elementary indicators are required, able to discriminate cases on the continuum, referring to the characteristic, in points that are different from each other. In other words, elementary indicators have to contribute to the description of the measured characteristic in different (cumulative) manners. The cumulative models are able to evaluate and verify the capacity of the selected elementary indicators to respect this cumulative requirement. In particular, cumulative models are based upon the following requirements:

- *unidimensionality*: the group of selected elementary indicators refers to a single conceptual dimension,
- each indicator has a *differentiated relationship* with the conceptual dimension. The consequences is the

- *absence of compensability* among the selected indicators. This requirement can be operationalized in terms of *graduality/scalability*. This means that indicators should be selected so that they turn out to be discriminant at different points of the same conceptual dimension. In other words, it should be possible to arrange the selected indicators in different points according to an increasing level of intensity. The indicators can also show some partial overlapping meaning that allows graduality in measurement,*homogeneity* of scaling techniques, that is the same classifications should be applied for whole group of indicators,
- the group of indicators should be *exhaustive*, that is should cover all the variability allowing a global evaluation.

Historically, Louis Thurstone (1927, 1959) was the first researcher engaged in the creation a continuum with a increasing intensity concerning a certain characteristic by using the judgments expressed by a group of "judges" (Arcuri & Flores D'Arcais, 1974; McIver & Carmines; 1979; Torgerson, 1958) and his approach is often applied in order to obtain differential scales.
In particular, Thurstone was mainly concerned with the fundamental problem of how psychological stimuli could be measured and compared with one another.

If a researcher aims at identifying the "weight" of each of a set of objects (non-physical) – such as, occupations with reference to the characteristic of prestige – the task turns out to be problematic since no reference scale is available. In this case, the process of ordering the objects by their relative prestige can be accomplished by multiple subjective judgments that could collected through two different procedures: (a) each of a group of individuals is asked to arrange the objects according to a given criterion (e.g. "prestige": from the most prestigious to the less prestigious); (b) the objects can be presented in all possible pairs to each individual that points out the one that in the dyad better represents the criterion (possesses the characteristic at the highest – or lowest – level, e.g. the most prestigious occupation between two).

The model that he proposed is based upon a fundamental assumption, the *law of comparative judgments*. According to this law, each object (occupation) submitted to the individual judgment arises a response produced by a *discriminant process* referring to the considered attribute. This discriminant process is a theoretical construct and represents the evaluation expressed by an individual in comparing two objects with reference to the attribute.
We can assume for each object/stimulus and each attribute the existence of several *discriminant processes*. This means that the value of the discriminant process as a result of repeated evaluations of each object can show variations related to the existence of the error of measurement. This variability assumes the existence of a distribution of the discriminant processes. The distribution of the discriminant processes is assumed to be normal, described by two parameters, mean and standard deviation. The most frequently occurring response represents the *modal discriminant process* that defines the scale value of the object by which each object can be located along the continuum.
The basic assumption underlying the law of comparative judgment is that the degree to which any two objects can be discriminated is a direct function of the difference in their status as regards the attribute in question. If the great part of the respondents judges object A different from object B with reference to the continuum, the placement of objects on the continuum should reflect the degree to which respondents can discriminate among the perceived characteristic of the various objects.
The greater the distance between object A and object B on the continuum, the greater the proportion of respondents that have agreed that object A differs from object B. On the contrary, the smaller the distance between object A and object B on the continuum, the more confusion exists about the relative difference between the two objects with reference to the considered characteristic (McIver & Carmines; 1979; Thurstone, 1927, 1959; Torgerson, 1958).
Scales created by this method are called *Thurstone scales* or *differential scales*. Many analytical versions exist according to the experimental model adopted (assumptions) and on the number of cases and the number of objects involved.
Values, calculated through the application of particular and simple analytical procedure,[3] allow defined elements to be placed on the continuum and can be considered in terms of group subjective weights.
The main problem shown by this approach concerns the theoretically possibility to meet its fundamental assumptions, e.g. uni-dimensionality of the psychological continuum (McIver & Carmines, 1979).
The approach needs particular care from the applicative point of view, especially with reference to choice of (i) the objects that should be involved and that should share the same continuum (ii) the technique by which the objects should be showed be shown and evaluated by the respondents objects'. With reference to this, it should be considered that the paired comparison technique should not be applied with a high number of objects that could make the respondents' task too heavy, in terms of both time and required attention (Arcuri & Flores D'Arcais, 1974). Some solutions have been studied in order to make respondent's task lighter and easier.

[3] The actual analytical procedure to be applied in case of both comparison and ranking data is briefly illustrated in appendix D (McIver & Carmines; 1979; Thurstone, 1927, 1959; Torgerson, 1958).

Later on, two different cumulative models have been conceived and defined. These cumulative models refer to different approaches in dealing with measurement error (non-systematic variation in scores or non-systematic variance or error variance) and consequently to different definitions concerning the response model (called *trace line*); these models are:

- Deterministic model: according to this approach, the non-systematic variation is not explicitly definable and is completely attributed to the cases' and indicators' position on the continuum representing the measured dimension. Consequently, the probability to obtain a certain score for a certain indicator can be 0 (*beta*) or 1 (*alfa*) in any point of the underlying continuum. The approach known as "Guttman approach" represents the most common version of deterministic model and found applications in subjective measurements (Guttman, 1945, 1947; McIver & Carmines, 1979; Torgerson, 1958). Multidimensional versions of this model were proposed (Borg & Shye, 1995; Shye, 1985).
- Probabilistic model: according to this model, the random error can be defined as the probability to obtain a certain score. This approach, based upon the *Item Response Theory* (*IRT*) based upon the *Latent Trait Theory*, attributes the variation to both cases' characteristics (capacity, attitude, opinion, or others) and indicators' characteristics (difficulty or discriminant capacity). The obtained score represents a measure of the relationship between each case and each indicator. The relationship is formally described by the *Item Characteristic Curve* (*ICC*). Unidimensionality and local independence are the basic assumptions. The definition of mathematical-probabilistic models allowed subsequently statistical criteria to be defined in order to test goodness of fit (Andersen, 1972; Andersen, 1973; Andrich, 1988; Hambleton et al., 1991; Lord, 1952, 1974, 1980, 1984; Ludlow & Haley, 1995; McDonald, 1989; Rasch, 1960; Sijtsma & Molenaar, 2002; Swaminathan & Gifford, 1982, 1985, 1986; Torgerson, 1958).

4.3 Perceptual mapping

The terms *perceptual mapping* refer to several scaling models, which have the common goal to identify and represent the underlying dimensions the obtained scores (generally subjective reactions to submitted objects). In the past, these scaling models found wide applications in marketing research since perceptual maps allow mental structures to be identified.

These models require data represented by proximity (similarity or preference) matrices and are based upon analytical methods that in some cases found application as generic multivariate statistical analysis methods since they allow:

- dimensions underlying the obtained scores (expressed in terms of similarities or preferences concerning selected objects) to be identified,
- relative importance of each dimension to be tested,
- objects to be adequately represented in the geometrical space, defined by the dimensions identified from similarities or preferences data.

The approaches can be distinguished into two main groups:

- approaches requiring similarities data. *Multidimensional Scaling* belongs to this group and presents analytical techniques (metric and non-metric, for individual or aggregated data). The metric approach recalls factor analysis (Cox & Cox, 1994; Kruskal & Wish, 1978; Torgerson, 1958);
- approaches requiring preferences data. Preferences data are represented as geometrical relationships between points in an unidimensional or multidimensional space. The *unfolding* method represents the most known approach (Coombs, 1950; McIver & Carmines, 1979).

4.3.1 Multidimensional Scaling

Synthetically, the application of this approach is carried on through the following stages:

a. construction of the proximities matrix (e.g. similarities expressed by individuals concerning a group of objects);
b. definition of a distance model and identification of a spatial model in order to transform the obtained proximities into distances;
c. mapping the objects through an iterative procedure starting from a initial randomly defined configuration. The iterative procedure proceeds by:
    - o computing distances between the objects from the observed proximities through the models previously defined,
    - o comparing two matrices, containing respectively the observed proximities (δ) and the computed

distances (d), and evaluating the goodness of fit (indexes are provided: e.g. stress or alienation coefficients);[4]
  - o computing disparities between proximities and distances;

the iterative procedure will stop when comparison between distances and proximities will not show a goodness of fit better then the previous iteration;

d. computation of the coordinates for each object according to the adopted spatial model (matrix ***X***, in which each cell contains x_{ir} representing the coordinates for each object *i* for each dimension *r*) and graphical representation of the points;

e. analysis of the relationships between objects and interpretation of the obtained dimensions (eventually rotated).

4.3.2 Unfolding model

The unfolding approach is one of the models developed for the preferential choice data. It is aimed at representing subjects and objects (said "stimuli") in a common space – usually unidimensional – such that the relative distances between them reflect the psychological proximity of the objects to the individuals. The analytic approach, defined and introduced by Coombs (1950; McIver & Carmines, 1979), allows one preference scale (or more scales) to be obtained from the rankings of the objects made by the subjects.
The procedure requires the administration of a series of stimuli that have to be ordered by each subject according to a preference criterion. Each individual's preference ordering is called *I scale*.[5]
The basic assumption posed by the model states that one (or more) common latent attribute (referred to as joint scale or *J scale*) exists underlying the different observed preference orderings of a group of individuals. The underlying dimensions can be determined as a result of the identification of the *ideal point* of the scale on which the subject is placed. The goal is to verify whether the different individual *I* scales can be located in a single *J* scale.[6] If so, then we can reasonably conclude that the subjects employ a common criterion in evaluating the various stimuli. In the opposite case, two different possibilities exist:

- subjects employ multiple criteria in the evaluation of the stimuli,
- subjects respond to the stimuli in a personal way, in other words, a common underlying attribute does not exist.

Let us suppose that two subjects expressed their preferences with reference to five stimuli – *a, b, c, d, e* – and that the preferences could be represented on a single dimension. The process of evaluating the consistency of the individual *I* scales to be represented on a common *J* scale is called unfolding the *I* scales.
The following figure (McIver & Carmines, 1979) illustrates the process.
The vertical lines I_1 and I_2 represent the individual orderings of the two subjects, respectively *cbade* and *decba*, while the horizontal line represents the *J* scale.
We assume that the "strength" of preference expressed by each subject in a single dimension can be represented by a normal distribution. In this model, the more distant the object from the mean of preferences distribution, the less preferred the object.
If the axioms of distances (Maggino, 2005a) are acceptable, then the direction will not be involved in computing preferences. At this point, it is possible to proceed according to two different perspectives:

- <u>*Unfolding*</u>: according to this perspective, individual preference orderings (*I* scales) can be used in order to determine the *J* scale (strength of preference). The figure shows in which way portions of the I_1 and I_2 (*unfolding* lines) scales can be individuated in order to define the *J* scale. This scale preserves the essential integrity of the individual *I* scales in the sense that a particular stimulus is closer to the subject

---

[4] The most common coefficient is the following:

$$stress = \sqrt{\frac{\sum\left(\delta_{ij} - d_{ij}\right)^2}{\sum \delta_{ij}^2}}$$

where

δ_{ij} observed proximity between object *i* and object *j*

d_{ij} computed distance between object *i* and object *j*

[5] The *unfolding* input matrix is *two-mode two-ways*. The generic element a_{ij} represents the preference expressed by the *j*-th individual with reference to the *i*-th object. The model allows the two modes of the matrix to be represented in a single spatial representation: the *N* objects and the *m* individuals (*joint space analysis*).

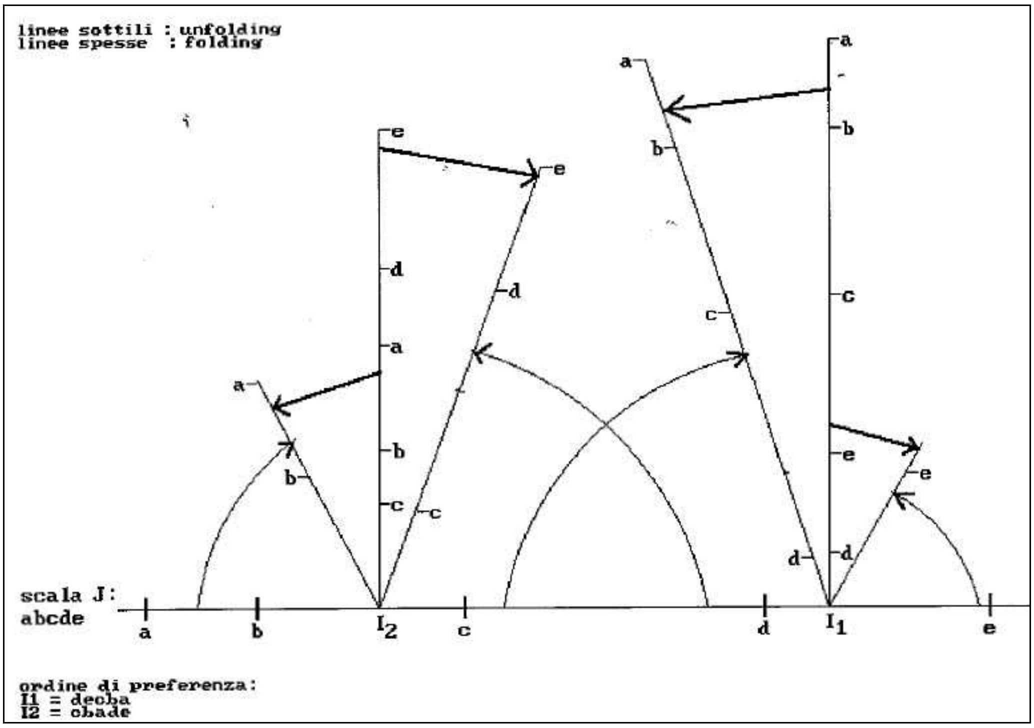
[6] With reference to this, the model distinguishes between:
- o <u>qualitative *J* scale</u>, represented by the simple order of the objects ((the distance between objects is unknown);
- o <u>quantitative *J* scale</u>, definable when distances between objects can be inferred from the order of the objects.

that is preferred to another. We can observe that according to I_1

- stimulus *c* is preferred to stimulus *b*: on the *J* scale, I_1 is closer to *c* than to *b*,
- stimulus *d* is preferred to stimulus *c*: on the *J* scale, I_1 is closer to *d* than to *c*.

The observed relation preference-distance can be observed also on the individual I_2 scale. Consequently, both the individual orderings can be *unfolded* on the same dimension.

- *Folding*: according to this perspective, the *J* scale can be used in order to draw the individual preference orderings (*I* scales). The figure shows in which way it is possible to individuate portions of the *J* scale (*folding* lines) that can be *folded* in order to re-arrange the individual I_1 and I_2 scales (orderings). This can be done by folding the *J* scale in relation to the ideal point representing each individual.



The arrows depicted in the figure help in identifying both the procedures: flat arrows are related to the *unfolding* procedure while curved arrows refer to the *folding* procedure.
Generally few individual scales (*I*) are employed given that the model application turns out to be more complex in presence of a great number of *I* scales (McIver & Carmines, 1979).

Multi-dimensional model. As seen, the *unfolding* approach is aimed to represent – on a single metric continuum – both stimuli and subjects from preferences expressed by a group of subjects. This approach assumes that subjects employ a common criterion in expressing the preferences with reference to the stimuli.
Some Coombs's scholars have extended the model to higher dimensions, applicable when the preferences are supposed to be expressed by respondents according to different criteria. The theoretical approach remains the same even if the geometric structure turns out to be more complex. The goal is to place the points regarding both the objects and the respondents in a *R*-dimensional space by using the distances, Euclidean or not.
Let us suppose that the objects are represented by candidates fro political elections and that the respondents are voters asked to rank the candidates with respect preferences. If "ideology" should be the unique preference criterion used by respondents in the evaluating process, then the preferences could be represented in an uni-dimensional space. on the contrary, if the voters evaluate the candidates according to also other characteristics (professional, personal, and so on), a multi-dimensional space should be identified in order to represent all the preferences.
The application of the multi-dimensional version of the model is made problematic by the difficulty to develop consistent goodness-of-fit algorithms. This difficult arises because in order to estimate a big number of information ($n*m$ matrix concerning the subjects' points co-ordinates and $k*m$ matrix of objects' points co-ordinates) a small number of information ($n*k$ matrix) is used. It follows that many points configurations are obtainable and are able to fit data acceptably. Consequently, the multi-dimensional approach should be carefully considered because the possibility exists to obtain degenerate solutions (local minimum).

4.4 Conjoint model

Conjoint measurement is an axiomatic theory of measurement that defines the conditions under which there exist measurement scales for two or more variables that jointly define a common scale under an additive composition rule (Luce & Tukey, 1964). This theory became the basis for a group of related numerical techniques for fitting additive models, called *conjoint analysis* (Green and Rao, 1971), known also as *multi-attribute compositional model* or *stated preference analysis*.

It was originated in the ambit of quantitative psychology and has found applications in many research fields, like *marketing research* or operational research. More recently, conjoint analysis methodology found different application in the field of designing experiments (Louviere, 1991).

Conjoint analysis is used specifically to understand how respondents develop preferences for certain objects (products, services, ideas, ambits and so on). It is based on the simple premise that individuals evaluate the value of an object (real or hypothetical) by combining separate amounts of value provided by each objects' attribute.

The goal is to determine which combination of attributes is that preferred by the individual (Hair, 1998; Louviere, 1988; Malhotra, 1996).[7]

Utility represents the conceptual basis for measuring value in conjoint analysis. It is a subjective judgment of preference unique to each individual. In conjoint analysis, utility is assumed to be based on the value placed on each of the values of the attributes and expressed in a relationship reflecting the manner in which the utility is formulated for any combination of attributes. We might sum the utility values associated with each feature of an object to arrive at an overall utility. Then we would assume that objects with higher utility values are more preferred and have a better chance of choice.

Conjoint analysis is unique among multivariate methods in that the researcher first constructs a set of real or hypothetical objects by combining selected values of each attribute. These combinations are then presented to respondents, who provide only their overall evaluations. As the researcher constructs the hypothetical objects in a specific manner, the influence of each attribute and each value of each attribute on the utility judgment of a respondent can be determined from the respondents' overall ratings.

Procedure. The researcher must identify the *factors* describing the specific object of interest, and then the *levels* values defining each factor.

Next, different configurations of the object are identified by combining different values (levels) for each factor. Each combination is named *scenario*.

Next, a group of respondents is asked to evaluate and rank alternative the scenarios according to a given criterion. The evaluation is expressed according to one of the following approaches:

- *ranking*: respondent ranks scenarios in order of preference,
- *rating*: respondent assigns to each scenario a level of preference expressed on a rating scale.

If the researcher built the scenarios by creating specific and appropriate factor-level combinations, the analysis of the expressed preferences allow the criteria of preference used to be identified and the subjective structure of preference to be understood.

In particular, the purpose of the analysis is – through a de-compositional process – that to determine

- the importance and the weight of each factor in the total subjective decision,
- how much each level of each factor has influenced the total preference (utility).

The *total worth*, expressed by a respondent with regard to an object, is formed of partial values (*part-worth*) relating to each level for each factor. The conjoint model can be formalized as following:

$$total \cdot worth = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(part - worth_{ij} \right)$$

where

m number of factors

n number of levels for each factor (value that changes for each factor).

Estimates of part-worths allow the respondent's preference for any combination of factors to be assessed.

[7] Since the mid of the Seventies, conjoint analysis has attracted considerable attention as a method that portrays consumers' decisions realistically as trade-offs among multi-attribute products or services. Conjoint analysis gained widespread acceptance and use in many industries. During the 1990s, the application of conjoint analysis increased even further, spreading to many fields of study. Marketing's widespread utilization of conjoint in new product development for consumers led to its adoption in many other areas.

At the same time the development of alternative methods of constructing the choice tasks for consumers and estimating the conjoint models was observed.

Accelerated use of conjoint analysis has coincided with the widespread introduction of computer programs that integrate the entire process, from generating the combinations of independent variable values to be evaluated to creating choice simulators for predicting consumer choices across a wide number of alternative product and service formulations.

Conjoint analysis is best suited for understanding consumers' reactions to and evaluations of predetermined attribute combinations that represent potential products or services. While maintaining a high degree of realism, it provides the researcher with insight into the composition of consumer preferences.

The preference structure could reveal which is/are the factor/s determining the total utility and the final choice. Value of an extreme or infeasible level should be deleted from the analysis or the importance values should be reduced to reflect only the range of feasible levels.
The analysis can be performed at both individual and group level. In particular, the choices expressed by a group of subjects can be combined in order to represent a "competitive" ambient.
This approach is considered *compensatory* and consequently requires a careful evaluation of its applicability.

Statistical characteristics of the model. Conjoint analysis presents the following main characteristics (Hair et al., 1998):

- o Decompositional model. Conjoint analysis *decompose* the total respondent's preference with reference to the object. Definition of the objects is carried out through a process aimed at specifying a set of attributes (factors) and a group of values (levels). Different combinations of levels regarding the identified attributes define different objects. The respondent is asked to express preference with regard the objects. Once given, the preference is decomposed to determine the value (importance) of each attribute by relating the known attributes of the object (which become the independent variables) to the evaluation (dependent variable).
- o Linear model. Conjoint analysis employs a *variate*, a linear combination of effects of the independent variables (factors) on the dependent variable (subject's choice). Both the independent variables (factors) and their values (levels) are specified, while the dependent measure is provided by the respondent. The specified levels are then used by conjoint analysis to decompose the respondent's response into effects for each level (much as is done in regression analysis for each independent variable). In this perspective, the project design represents a critical step in view of a good success of the study. If a variable or effect is not anticipated in the research design, then it will be not available for the analysis. For this reason, the researcher may be tempted to include a number of variables that might be relevant. On the other side, conjoint analysis is limited in the number of variables that can be included (the researcher cannot simply add new questions to compensate a weak conceptualisation of the problem). The goal is to develop a predictive model.
- o Testing and estimation of the model at individual level. The originality of this approach is mainly in that it can be carried out at the individual level. In other words, the researcher generates a separate model for predicting preference for each respondent. In conjoint analysis, however, estimates can be made for the individual (disaggregate) or groups of individuals (aggregate). At disaggregate level, each respondent rates enough stimuli for the analysis to be performed separately for each person. Predictive accuracy is calculated for each person. The individual results can be aggregated to portray an overall model as well. At aggregate level, the researcher is interested to perform the estimation of parth-worths for the group of respondents as a whole. Aggregate analysis can provide (i) a mean for reducing the data collection task through more complex designs, (ii) methods for estimating interactions, and (iii) greater statistical efficiency by using more observations in the estimation. In selecting between aggregate and disaggregate conjoint analysis, the researcher must balance the benefits gained by aggregate methods versus insights provided by the separate models obtained by disaggregate models.
- o Flexibility. Conjoint analysis is a quite flexible approach, since it allows:
  - metric and non-metric variables to be employed,
  - categorical variables to be employed as predictive variables,
  - separate prediction to be made for the effects of each level of the independent variable without assuming the correlation between them.
  - non-linear relationships to be easily handled. This is true also for complex curvilinear, in which one value is positive, the next negative, the third positive again, and so on.

5. *Methodological challenges in measuring subjective well-being*

By concluding this work, two challenges in measuring subjective well-being could be pointed out
1. Measures' assessment
2. measurement requirements for comparative research

(1)

Concerning the first issue, Zumbo (2009, Zumbo & Forer, in press) proposes a multi-level framework for validity, using individual measurement results at various levels of a complex ecologically rich system, moving from individuals to aggregates.
In particular, he focuses on what happens, when the individual measures are aggregated and inferences are made at a different (higher) level in the system. In these cases inferences at a higher aggregate level (neighborhood, province, region, nation) may carry with secondary dimensions that may contaminate or confound the inference (form of "fallacies").
Zumbo and Forer (in press) note that:
- any inferences from the individual level may not hold in the same way at higher (or lower) levels of aggregation
- systematic and coherent evidence (validation evidence) needs to assembled to support the inferences at the various levels
- the level of validation evidence needs to be <u>*in line*</u> with the level of inferences
- individual level validity evidence (which is what is traditionally done in validation research, e.g. criterion validity) does not provide sufficient validity evidence for inferences at higher levels in the system; and may actually be misleading because it may miss invalidity at the aggregate level.

By summarizing, applying traditional individual differences validation methods (e.g., correlation with another wellbeing measure, or even cognitive response models) are insufficient evidence for support multi-level validation inferences like those often used in wellbeing research.
In fact, these individual differences validation methods are susceptible to the cross-level inferential fallacies such as the (reverse) ecological fallacy or atomistic fallacy.
Multi-level validation arises when one has a multi-level construct; that is, an individual level measure (or assessment) and aggregating it to make inferences at a higher level.
Historically, multi-level constructs have not been an issue in measurement and validation because measurement has been immersed in and emerged from an individual differences psychological or sociological school of thought.
In Zumbo's proposal, multi-level validation research might include, for example, the following issues:
- Is the aggregate score reflecting differences between measurement units (at that level) such as national differences in wellbeing?
- To what extent might we be measuring, unintentionally, other important constructs <u>at the aggregate level</u> that are not meant to be included in our measures (at that level), such as, construct irrelevant variance like neighborhood effects, or regional effects, or is our measure of wellbeing mostly a restatement of gross domestic product (or other such economic indicators)?
- Also a matter of determining what is that is (and is not) be measured by that aggregate variable.

(2)

In recent Saris & Gallhofer's work (2007), problems of **cross-cultural comparative research** are discussed.
It is well known that measurement error has strong effects on results of research. Therefore, when the effects of measurement error differ in the individual countries, comparisons across countries become quite challenging.
Two types of comparisons are most frequently made:
- comparison of *means* and
- comparison of *relationships* of different variables across countries.

Often comparisons based on single questions or on composite scores of the latent variables are made.
The author adds to this the comparisons based on latent variables.

The problem of such comparisons is that one can compare the results across different countries only if in fact the data are comparable, that is, if the measures used in the different countries have the same meaning.
This topic is studied under the heading of **functional equivalence** or **invariance of measures** in different countries.
Saris's work concentrates on the procedures to determine equivalence of measurement instruments.
The measurement requirements for comparative research (reflective indicators) are:

- **configural invariance** → the same standard factor analysis model should hold for all different groups
- **metric invariance** → the equality of the loadings
- **scalar invariance** → the intercepts should also be equal in the different groups

These requirements are **too strict**.
There are **two reasons** for this.

a) A response model can be specified that makes a distinction between
   - the *interpretation of the questions* →<u>cognitive part</u> of the model. It should be the same across groups because otherwise people have different ideas about the concepts of interest.
   - the *response process* → the <u>measurement part</u> of the model. Any differences observed here are less fundamental.

   The differences in this measurement process can be separately estimate and correct for these differences. Suggestion: the above mentioned requirements for comparative research should hold after correction for measurement errors
b) Significant differences across countries test are done for parameters of single indicators while these indicators are combined to an index. Therefore, a significant deviation of one indicator across countries in a set of other indicators may have only a very minimal effect on the total score for the index and the deviation may be rather irrelevant evaluating the index as a whole.

References

Ajzen, I. and M. Fishbein (1980) *Understanding Attitudes and Predicting Social Behavior*, Englewood Cliffs, NJ: Prentice-Hall,Inc.

Amaturo E. (1989) *Analyse des données e analisi dei dati nelle scienze sociali*, Centro Scientifico Editore, Torino.

Andersen E.B. (1972) "The Numerical Solution of a Set of Conditional Estimation Equations", in *Journal of the Royal Statistical Society*, Serie B, 34.

Andersen E.B. (1973) "A Goodness of Fit Test for the Rasch Model" in *Psychometrika*, 38.

Andrews, F.M. and S.B. Withey (1976) *Social Indicators of Well-being: Americans' Perceptions of Life Quality*, Plenum Press, New York-London.

Andrich D. (1988) *Rasch Models for Measurement*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-068, Newbury Park, CA:Sage.

Arcuri L. and Flores D'Arcais G.B. (1974) *La misura degli atteggiamenti*, Martello – Giunti.

Aureli Cutillo E & E. Koch-Weser Ammassari (eds.) (1977) *Socializzare la statistica*, Università degli Studi "La Sapienza", Roma.

Bejar I.I. (1983) *Achievement testing. Recent Advances*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-036, Newbury Park, CA:Sage.

Biemer P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A., Sudman S. (1991) *Measurement Errors in Surveys*, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.

Borg I., Shye S. (1995) *Facet Theory. Form and Content*, Advanced Quantitative Techniques in the Social Sciences Series, Vol. 5, SAGE Pubblications, Thousand Oaks - London - New Delhi.

Bruschi A. (1999) *Metodologia delle scienze sociali*, Bruno Mondadori, Milano.

Campbell, D T & Fiske, D W. (1959) "Convergent and discriminant validation by the multitrait-multimethod matrix", *Psychological Bulletin*, 56:81-105

Campbell D.T., Russo M.J. (2001) *Social Measurement*, SAGE Publications, London.

Carmines E.C., Zeller R.A. (1992) *Reliability and Validity Assessment*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-017, Newbury Park, CA:Sage.

Coombs C.H. (1950) "Psychological scaling without a unit of measurement," *Psychological Review*, 57, 148-158.

Coombs C.H. (1953) "Theory and Methods of Social Measurements" in L. Festinger and D. Katz (eds.) *Research Methods in the Behavioral Sciences*, New York: Dryden Press.

Coombs C.H. (1964) *A theory of Data*, Ann Arbor, MI: Mathesis.

Cox T.F., Cox M.A.A. (1994) *Multidimensional scaling*, Chapman & Hall, London.

Cronbach L. J., P. E. Meehl (1955) "Construct validity in psychological tests", *Psychological Bulletin*, *52*, 281-302. http://psychclassics.yorku.ca/Cronbach/construct.htm#f1.

Dawes, John (2008), "Do Data Characteristics Change According to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales," International Journal of Market Research, 50 (1), 61-77.

Delli Zotti G. (1985) "Tipologia delle matrici utilizzate nella ricerca sociale" in *Rassegna Italiana di Sociologia*, Anno XXVI, n. 2.

De Vellis R. (1991) *Scale devolopment. Theory and Applications*, Applied Social Research Methods Series, vol. 26, SAGE Publications, London.

Fishbein, M. and I. Ajzen, (1975) *Belief, Attitude, Intention, and Behavior: An Introduction to Theory and Research.* Reading, MA: Addison-Wesley. http://www.people.umass.edu/aizen/f&a1975.html

Flament C. (1976) *L'analyse booléenne de questionnaire*, Mouton Éditeur, Paris, La Haye.

Ghiselli E.E. (1964) *Theory of Psychological Measurement*, McGraw-Hill, New York-London.

Green P.E., Rao V.R. (1971) "Conjoint measurement of quantifying judgmental data", *Journal of Marketing Research*, vol. 8, 355-363.

Groves, R.M. (1989) *Survey Errors and Survey Costs*, John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore.

Guttman L. (1945) "Questions and answers about scale analysis", Research Branch, Information and Education Division, Army Service Forces, Report D-2.

Guttman L. (1947) "On Festinger's Evaluation of Scale Analysis", in *Psychological Bulletin*, 44.

Hair J.F. jr., Anderson R.E., Tatham R.L., Black W.C. (1998, 5th ed.) *Multivariate Data Analysis*, Prentice-Hall Inc., Upper Saddle River, New Jersey.

Hambleton R.K., Swaminathan H., Rogers H.J. (1991) *Fundamentals of item response theory*, Measurement Methods for the Social Sciences series, vol. 2, SAGE Publications, London.

Hughes C. & Wang B. (1996) “Attempts to conceptualise and measure quality of life”, in R. L. Schalock (ed.), *Quality of Life, Vol. 1: Conceptualisation and Measurement*, American Association on Mental Retardation, Washington, DC., 51–62.

Jacoby W.G. (1991) *Data Theory and Dimensional Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-078, Newbury Park, CA:Sage.

Kruskal J.B., Wish M. (1978) *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-011, Newbury Park, CA:Sage.

Laveault D., Zumbo B.D., Gessaroli M.E., Boss M.W. (eds.) (1994) *Modern Theories of Measurement: Problems and Issues*, Edumetrics Research Group, University of Ottawa, Ottawa, Canada.

Likert, R. (1932), "A Technique for the Measurement of Attitudes", *Archives of Psychology* 140: 1–55.

Lodge M. (1981) *Magnitude Scaling. Quantitative Measurement of Opinion*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-025, Newbury Park, CA:Sage.

Lord F.M. (1952) *A Theory of Test Score* (Psychometric Monograph N.7), Iowa City, IA: Psychometric Society.

Lord F.M. (1974) "Estimation of Latent Ability and Item Parameters when there are Omitted Responses" in *Psychometrika*, 39.

Lord F.M. (1980) *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale, NJ: Lawrence Erlbaum.

Lord F.M. (1984) "Standard Errors of Measurement at Different Ability Levels" in *Journal of Educational Measurement*, 21.

Lord F.M., M.R. Novick (1968) *Statistical Theories of Mental Test Scores*, Reading, MA: Addinson-Wesley.

Louviere J.J. (1988) *Analyzing decision making: metric conjoint analysis*, Sage, Newbury Park

Louviere J.J. (1991) “Experimental choice analysis: introduction and review”, *Handbook of Marketing Research*, Oxford: Blackwell Publisher.

Lozano L.M., García-Cueto E., and Muñiz J. (2008) Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales, *Methodology* 2008; Vol. 4(2):73–79 Hogrefe & Huber Publishers

Luce R., Tukey J.W. (1964) “Simultaneous conjoint measurement: a new type of fundamental measurement”, *Journal of Mathematics Psychology*, vol. 1, 1-27.

Ludlow L.H., Haley S.M. (1995) "Rasch Model Logits: Interpretation, Use and Trasformation" in *Educational and Psychological Measurement*, Sage Periodicals Press, London, Vol. 55, N. 6, pp. 967-975.

Maggino F. (2003) *Method effect in the measurement of subjective dimensions*, Firenze University Press, Archivio E-Prints, Firenze.

Maggino F. (2007) *La rilevazione e l’analisi statistica del dato soggettivo*, Firenze University Press, Firenze.

Maggino F. (2005a) *L’analisi dei dati nell’indagine statistica*, Firenze University Press, Firenze.

Maggino F. (2005b) *The Importance of Quality-Of-Life Dimensions in Citizens’ Preferences: An Experimental Application of Conjoint Analysis*, Firenze University Press, Archivio E-Prints, Firenze.

Maggino F. & T. Mola, (2007) *Il differenziale semantico per la misura degli atteggiamenti: costruzione, applicazione e analisi. Presentazione di uno studio*, Firenze University Press, Archivio E-Prints, Firenze.

Malhotra N.K. (1996) *Marketing research: an applied orientation*, Prentice-Hall International, Inc., Englewood Cliffs, New Jersey

Marradi A. (1980) *Concetti e metodi per la ricerca sociale*, La Giuntina, Firenze.

Marradi A. (1990) “Fedeltà di un dato, affidabilità di una definizione”, in *Rassegna Italiana di Sociologia*, a. XXXI, n. 1, gennaio-marzo.

Maslow A. H. (1954) *Motivation and Personality*, Harper and Brothers, New York.

McDonald R.P. (1989) "Future Directions for Item Response Theory", *International Journal of Educational Research*, 13(2).

McIver J.P., Carmines E.G. (1979) *Unidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-024, Newbury Park, CA:Sage.

McKeown B. and Thomas D. (1988) *Q Methodology*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-066, Newbury Park, CA:Sage.

Michalos A. (1992) “Use and Abuses of Social Indicators” in *Sinet*, n. 32.

Netemeyer R.G., Bearden W.O., Sharma S. (2003) *Scaling Procedures: Issues and applications*, Sage Publications, Thousand Oaks.

Nunnally J.C. (1978) *Psychometric theory*, McGraw-Hill, New York - London.

Rasch G. (1960) *Probabilistic Models for Some Intelligence and Attainment Tests*, Copenhagen: Danish Institute for Educational Research.

Saris W.E. and I.N.Gallhofer (2007) *Design, evaluation and analysis of questionnaires for survey research*, New York: Wiley-Interscience.

Sawtooth Software (2004) *The MaxDiff/Web v.6.0 Technical Paper* Technical Paper available at http://www.sawtoothsoftware.com/download/techpap/maxdifftech.pdf.

Schultz W. (2000) *Explaining Quality of Life – The Controversy between Objective and Subjective Variables*, EuReporting Working Paper No. 10, Paul Lazarsfeld-Gesellschaft für Sozialforschung (PLG).

Siegel S. and N.J.Castellan j. (1952, ed. it. 1992) *Statistica non parametrica*, Mc-Graw-Hill Italia.
Sijtsma K., Molenaar I.W. (2002) *Introduction to Nonparametric Item Response Theory*, Measurement Methods for the Social Sciences series, vol. 5, SAGE Publications, London.
Spector P.E. (1990) *Research Designs*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-023, Newbury Park, CA:Sage.
Spector P.E. (1992) *Summated Rating Scale Construction. An Introduction*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-082, Newbury Park, CA:Sage.
Stevens, S.S. (1946) “On the theory of scales of measurement”, in *Science*, 103, 677-680.
Stevens, S.S. (1951) “Mathematics, measurement and psychophysics”, in S.S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.
Stevens, S.S. (1957) “On the psycophysical law”, in *Psychological Review*, 64, 153-181.
Swaminathan H., J.A. Gifford (1982) "Bayesian Estimation in the Rasch Model" in *Journal of Educational Statistics*, 7.
Swaminathan H., J.A. Gifford (1985) "Bayesian Estimation in the Two-Parameter Logistic Model" in *Psychometrika*, 50.
Swaminathan H., J.A. Gifford (1986) "Bayesian Estimation in the Three-Parameter Logistic Model" in *Psychometrika*, 51.
Thompson B. ed. (2003) *Score Reliability. Contemporary thinking on reliability issues*, SAGE Publications, London.
Thurstone, L.L. (1927). *A law of comparative judgement*. Psychological Review, 34, 278-286.
Thurstone, L.L. (1959) *The Measurement of Values*. Chicago, The University of Chicago Press.
Torgerson W.S. (1958) *Theory and Methods of Scaling*, John Wiley & Sons, Inc., New York, London, Sydney.
Traub R.E. (1994) *Reliability for the Social Sciences - Theory and Applications*, Measurement Methods for the Social Sciences series, vol. 3, SAGE Publications, London.
Trochim, W.M. (1999, 2nd ed, 2000) *The Research Methods Knowledge Base*, Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/index.php
Velleman, P. F. and Wilkinson, L. (1993) «Nominal, ordinal, interval, and ratio typologies are misleading” in *The American Statistician*, 47(1), 65-72,
at URL http://www.spss.com/research/wilkinson/Publications/Stevens.pdf
Weller S.C., Romney A.K. (1990) *Metric Scaling. Correspondence Analysis*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-075, Newbury Park, CA:Sage.
Zumbo B. (2009) “Validity as Contextualized and Pragmatic Explanation, and Its Implications for Validation Practice” in Robert W. Lissitz (Ed.) *The Concept of Validity: Revisions, New Directions and Applications*, IAP - Information Age Publishing, Inc.: Charlotte, NC.
Zumbo, B. D., & Forer, B. (in press) “Testing and Measurement from a Multilevel View: Psychometrics and Validation”, In James Bovaird, Kurt Geisinger, & Chad Buckendahl (Eds) *High Stakes Testing in Education - Science and Practice in K-12 Settings [Festschrift to Barbara Plake]*. American Psychological Association Press, Washington, D.C..